Big data approaches to investigating Child Mental Health disorder outcomes

Downs, Jonathan Muir

Awarding institution:
King's College London

BIG DATA APPROACHES TO INVESTIGATING CHILD MENTAL HEALTH DISORDER OUTCOMES

JOHNNY DOWNS

Thesis submitted for the degree of Doctor of Philosophy

September 2017

Department of Psychological Medicine

Institute of Psychiatry, Psychology & Neuroscience

King's College London

ABSTRACT

Background: In the UK, administrative data resources continue to expand across publically funded youth-orientated health, education and social services. Despite attempts to capture these data in structured formats, which are more accessible for analysis, most health information is stored as free text entry in electronic records. Big data techniques which combine large scale data linkage and automatic information extraction from free text, using Natural Language Processing (NLP), have considerable potential for enhancing the depth of information available from routinely collected public service data. There are a very limited number of published studies which have applied these big data techniques to answer questions relevant to child and adolescent psychiatry.

Methods: This thesis examined original and clinically relevant research questions using data from routinely collected electronic health records, enriched by NLP and linkages to external data sources. Five related studies were performed all using data obtained from the SLaM BRC Case Record Information Search (CRIS) extracted using a NLP approaches, with two studies using external linkages with routinely collected national electronic datasets (NHS Hospital Episode Statistics and DfE National Pupil Database, NPD).

Results: Using these data resources, I provide empirical support for the hypothesis that neurodevelopmental comorbidities increase children and adolescents' risk for potentially more harmful treatments, greater treatment complexity and worse clinical outcomes. The NLP methods employed overcame limitations of structured data extraction, providing better assessment of a diverse range of symptom types, severity and related impairments, including suicidal risk, negative symptoms, antipsychotic treatment failure, and self-harm. External data linkages with the NPD enabled population level analyses by nesting clinical samples within their source population. NPD linkage also permitted the inclusion of education performance data, which were not routinely available within electronic health records.

Conclusion: The thesis illustrates how the legal, governance and technical challenges were surmountable to enable linkage between NHS and Department for Education public service data. Also, it demonstrated that NLP and data linkages of electronic health records, have a clear role in clinical epidemiological studies of child and adolescent mental health. These tools, combined with the continued digitisation of public service activity, can unlock huge and detailed data resources for population-based analyses. However, current approaches have deficiencies, including limitations in accuracy, construct validity, and restrictions in the data available, providing challenges for future research.

ACKNOWLEDGEMENTS

I have been extremely fortunate to have Dr Richard Hayes, Professor Tamsin Ford and Professor Matthew Hotopf as supervisors. It was Tamsin and Matthew's generosity and guidance, which kick-started this research, turning a set of disparate ideas into a MRC funded research project. Richard was the catalyst to finally undertaking the PhD, and I am very grateful for all his support over the course of the last 4 years. He has been a tremendous mentor, balancing my impetuous nature, and kindly reinforcing what I need to do to move from a research idea, to a grant application, and onto published articles. Richard, Tamsin and Matthew continue to inspire me, and fuel my post-doctoral aspirations for child and adolescent mental health research. This PhD has been the most enjoyable period of my working life, which I believe is due to them, and the research environments they have created.

I would also like to thank Professor Robert Stewart and Matthew Broadbent who have had a considerable influence on the work within this thesis. Over the course of my PhD, they have developed the BRC Nucleus into a lively, multi-disciplinary workspace; it has provided me with rich opportunities to build valuable friendships and collaborations. Of these, a special mention must go to Dr Laura Pina, who was wonderful to work with, and whose drive and enthusiasm for understanding what might improve lives of children with early onset psychosis, was infectious. Thank you to Professor Ruth Gilbert, who has kept me alert to some of the methodological flaws commonly hidden within data linkage research, but also introduced me to key researchers in the child health and education policy arena, and generously supported me as an early career researcher. Thank you to Richard White, the National Pupil Database Team, and Dr Murat Soncul for helping me drive through the innovative linkage work. I am very grateful to Professor Emily Simonoff, Dr James MacCabe and Dr Rashmi Patel for all their contributions to the research contained in this thesis, and for all their help in shaping my future research ideas. Similarly, I would like to thank Dr Kate Polling, Dr Sumithra Velupillai, Dr Rina Dutta, and Dr George Gkotsis for their important contributions to the thesis and being great colleagues. Thanks to Dr Omer Moghraby for supporting me in keeping my hand in clinically, and his enthusiasm for applying research into clinical practice. I am incredibly grateful for all the practical and moral support that Megan Pritchard, Leo Koeser, Clare Taylor, Craig Colling, Anna Kolliakou, David Chandran, Ryan Little, Jyoti Jyoti, Debbie Cummings, Amelia Jewell and Hitesh Shetty have given me whilst I've been at the BRC.

Poppy, thanks for keeping it real. None of this work would have been possible without your love.

TABLE OF CONTENTS

APPENDICES

A Cohort table describing participant entry into chapter 8 study

B Letters of approval from national research governance committees

i. Oxford C Research Ethics Committee, reference 08/H0606/71+5 (Chapters 2-8)
ii. NHS Health Research Authority Confidentiality Advisory Group, reference: CAG 9-08(a)/2014 (Chapters 6-8)
iii. NHS Health Research Authority Confidentiality Advisory Group, reference: ECC 3-04(f)/2011 (Chapters 6-8)
iv. South London and Maudsley NHS Trust and Department for Education Memorandum of Understanding for sharing data (Chapters 6-8)

.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

aH.R	adjusted Hazard Ratio
APC	Admitted Patient Care
ASD	Autism Spectrum Disorder
CAG	Confidential Advisory Group
CAMHS	Child and Adolescent Mental Health Services
CDLS	Confidential Data Linkage Service
CRIS	Clinical Record Interactive Search
DfE	Department for Education
ED	Emergency Department
EHR	Electronic Health Records
EOP	Early Onset Psychosis
FSM	Free School Meals
HES	Hospital Episode Statistics
ICD	International Classification of Diseases
KS	Key Stage
MTF	Multiple Treatment Failure
NHS	National Health Service
NLP	Natural Language Processing
NPD	National Pupil Database
NPV	Negative Predictive Value
NS	Negative Symptoms
NSR	Not suicide related
aO.R	adjusted Odds Ratio
PPV	Positive Predictive Value
SDQ	Strengths and Difficulties Questionnaire
SR-Pos	Suicide related-Positive
SR-Neg	Suicide related-Negative
SEN	Special Educational Need
SLaM	South London and Maudsley NHS Foundation Trust
SQL	Structured Query Language
TP	True Positive

PUBLICATIONS

Primary publications arising from this thesis

Downs J, Dean H, Lechler S, Sears N, Patel R, Shetty H, Hotopf M, Ford T, Diaz-Caneja MD, Arango C, McCabe JH, Hayes RD, Pina-Camacho L. Negative symptoms in early-onset psychosis and their association with antipsychotic treatment failure (*Schizophrenia Bulletin*, under revision)

Ford T, Stewart R, Downs J. Surveillance, Case Registers and Big Data. In: Prince M, Stewart R, Ford T, Hotopf M, Das-Munshi J, eds. *Practical Psychiatric Epidemiology*, Second Edition UK. Oxford University Press (in press).

Downs J, Lechler S, Dean H, Sears N, Patel R, Shetty H, Simonoff E, Hotopf M, Ford T, Diaz-Caneja MD, Arango C, McCabe JH, Hayes RD, Pina-Camacho L. The association between co-morbid autism spectrum disorders and antipsychotic treatment failure in early-onset psychosis: a historical cohort study using electronic health records. *Journal of Clinical Psychiatry* (in press)

Downs J, Velupillai S, Gkotsis G, Holden R, Kikoler M, Dean H, Fernandes A, Dutta R. Detection of Suicidality in Adolescents with ASD: Developing a Natural Language Processing Approach for Use in Electronic Health Records. *Proceedings of the American Medical Informatics Association*. (in press)

Downs J, Gilbert R, Hayes RD, Hotopf M, Ford T. Linking up data to plan and improve mental health services for children in England. *Archives of Diseases in Childhood* 2017;102: 599-602

Downs J, Hotopf M, Ford T, Simonoff E, Stewart R, Shetty H, Jackson R, Hayes RD. Clinical predictors of antipsychotic use in children and adolescents with autism spectrum disorders: a historical open cohort study using electronic health records. *European Child and Adolescent Psychiatry* 2016; 25: 649-658

Secondary publications arising from this thesis

Velupillai, S, Hadlaczky, G, Baca-Garcia, E, Gorrell, GM, Werbelo, N, Nguyen, D, Patel, R, Leightley, D, Downs, J, Dutta R. Predicting and preventing suicidal behaviour: do risk assessment tools or data-driven approaches have a role? (*Journal of Mental Health*, under review)

Velupillai, S, Suominen, H, Liakata, M, Roberts, A, Shah, A, Morley, K, Osborn, D, Hayes, J., Stewart, R, Downs, J, Dutta, R. The Interplay of Evaluating Natural Language Processing Approaches and Clinical Outcomes Research (*Journal of Biomedical Informatics*, under review)

Ottisova L, Smith P, Stahl D, Shetty H, Downs J, Oram S. Psychological consequences of child trafficking: a historical cohort study of young survivors in contact with secondary mental health services (*PLoS One*, under revision)

Legge SE, Hamshere M, Hayes RD, Downs J, O'Donovan MC, Owen MJ, Walters JTR & MacCabe, JH. Reasons for discontinuing clozapine: A cohort study of patients commencing treatment. *Schizophrenia Research* 2016*;* 174: 113-119

Bogdanowicz KM, Stewart RJ, Chang CK, Downs J, Khondoker MD, Shetty H, Strang JS & Hayes RD. Identifying mortality risks in patients with opioid use disorder using brief screening assessment: Secondary mental health clinical records analysis. *Drug and Alcohol Dependence* 2016; 164: 82-88

Kadra G, Stewart R, Shetty H, Downs J, MacCabe JH, Taylor D & Hayes RD (2016) Predictors of long-term (≥6months) antipsychotic polypharmacy prescribing in secondary mental healthcare. *Schizophrenia Research* 2016; 174:106-12

Thompson JV, Clark JM, Legge SE, Kadra G, Downs J, Walters JT, ... MacCabe JH. Antipsychotic polypharmacy and augmentation strategies prior to clozapine initiation: A historical cohort study of 310 adults with treatment-resistant schizophrenic disorders. *Journal of psychopharmacology* 2016; 30: 436-443

Perera G, Broadbent M, Chang CK, Downs J et al Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record derived data resource. *BMJ Open* 2016; 6: 1-22 e008721

Lancefield K*, Randino A*, Downs J, Laurens K. Trajectories of childhood internalising and externalising psychopathology and psychotic-like experiences in adolescence: A prospective population-based cohort study. *Development and Psychopathology* 2016; 28: 527-536

Hayes, RD, Downs J, Chang CK. Shetty H et al. The Effect of Clozapine on Premature Mortality: An Assessment of Clinical Monitoring and Other Potential Confounders. *Schizophrenia Bulletin*, 2015; 41: 644-655.

CONTRIBUTION STATEMENT

I led on the study design, analysis and manuscript preparation for all studies described within the thesis. Data extraction from the Clinical Record Interactive Search (CRIS) database, and linked HES databases were completed in collaboration with the CRIS team. I wrote plain English specifications of the CRIS data extraction schedules and algorithms, with detailed instructions on the type and timing of the variables to be used. CRIS informaticians translated these into SQL extracts, and provided the raw data as flat data files.

The Natural Language Processing (NLP) applications described in chapters 2-8 were developed with data scientists based at King's College London, who were co-authors on submitted journals articles.

I wrote script for all derived variables post-extraction and analyses in STATA under supervision from my supervisors. I conducted all linked National Pupil Database extraction and analyses.

Manual clinical validation of the extracted data against the anonymised clinical record, was either performed by myself, or via clinical raters who I supervised, funded by an external grant from Foundation of Professional Services to Adolescents, awarded to me as principle investigator.

With exception of the main ethical approval governing the CRIS database, I led on gaining the ethical, legal and governance approvals to link and analyse all external sourced databases (HES and NPD). I conducted all the data analyses, wrote manuscripts (in collaboration with my supervisors and co-authors), submitted and published my research in peer-reviewed journals.

The work leading to this thesis was funded by a Medical Research Council Clinical Research Training Fellowship (MR/L017105/1)

ETHICS STATEMENT

All research conducted within the thesis has been conducted under the following ethical and governance approvals

- Oxford C Research Ethics Committee, reference 08/H0606/71+5 (Chapters 2-8)
- NHS Health Research Authority Confidentiality Advisory Group, reference: CAG 9-08(a)/2014 (Chapters 6-8)
- NHS Health Research Authority Confidentiality Advisory Group, reference: ECC 3-04(f)/2011 (Chapters 6-8)
- Department for Education Data Management Access Panel, reference: DR140613.01 (Chapters 6-8)

CHAPTER 1. INTRODUCTION

The contents of this chapter have contributed to the following:

Book chapter

Ford T, Stewart R, Downs J. Surveillance, Case Registers and Big Data. In: Prince M, Stewart R, Ford T, Hotopf M, Das-Munshi J, eds. *Practical Psychiatric Epidemiology*, Second Edition UK. Oxford University Press (in press).

Commissioned report

*Aschan L,*Downs J, Hotopf M**.** The Mental Health Landscape in England. NHS England. 2016

1.1 WHAT ARE BIG DATA?

Modern health and public services now collect huge amounts of electronic data. In 2011, the storage requirement for holding health and social care records for 9 million people, equivalent to the population of a very large city, was over 40 petabytes (40 x 107 gigabytes). By 2020, the annual rate of health data generation will be 40 times greater than it was in 2009.[1,2] There is no rigorous definition for Big Data, but at its simplest, Big Data refers to any electronic data that cannot be feasibly stored or processed by standard desktop computers or fit into a standard relational database.[3] The main features are high volume, variety and velocity: the '3 V's'.[2] Volume refers to the size of the data, where it is not uncommon for terabytes or petabytes to be available for analyses. Variety refers to the different types of data format (structured fields, free text, images, video, etc.) and multiple contributing sources, for example, health-related data could be derived from both social media and clinical notes. Velocity indicates the dynamic nature of data, where the volume and types of the data held are changing or evolving – for example, in a database derived from electronic medical records that updates in real time. 'Veracity' has been proposed as a fourth 'V', particularly in relation to social media data,[4] to underline the challenges in ascertaining the truthfulness of information recorded in these large-volume sources.

The term Big Data also concerns the tools used to interpret these complex structures. Over the years these have been referred to as data mining, analytics and, more recently, data science.[5] These terms describe the development and application of analytical techniques that can integrate and extract meaning from massive datasets. A common feature of these technologies is the capability to interrogate data using automated procedures, or algorithms, replacing the need for resource-intensive manual extraction or interpretation. Two promising Big Data approaches for healthcare research, and a focus of this thesis, are data linkage and Natural language processing (NLP). Data linkage provides the capacity to expand the types of variables beyond those usually collected by health systems, and analyse vast numbers of individual records held on separate clinical/non-clinical databases. NLP provides the capacity to detect a greater detail of clinical information, such as symptoms, medications and response to treatment from the free text, which can then be used in risk factors-outcome analyses.[6]

1.2 CURRENT CHALLENGES FOR CHILD & ADOLESCENT MENTAL HEALTH EPIDEMIOLOGY

Improving the prevention and treatment of mental disorders is one the greatest health challenges of the 21st century. In England, more than 850,000 children and 7 million adults have a mental disorder.[7,8] Mental disorders are the leading cause of disability in the UK, they represent 23% of the disease burden, and cost the UK £60 billion each year.[9] Most mental disorders are treatable.[10] England, as with many high income countries, offers a range of highly effective interventions, but access is often restricted and unequal across the population.[11,12] As it stands, up to 43% of children and 60% of adults with a psychiatric disorder will remain undiagnosed and untreated.[7,8]

Mental disorders are not a single entity, but represent a broad range of different symptoms and diagnoses, all on a spectrum of severity, from mildly impairing to life-threatening. Over half of adults with mental disorders first experienced signs and symptoms by the age of 14.[13] Childhood and adolescent mental disorders, if not properly addressed, can lead to significant adversity throughout adulthood.[14] Mental illness profoundly impacts children's access to learning and education - 17% of children with anxiety and depression are absent for a significant amount of school compared to 5% of children without mental disorders.[7] Each child with a mental disorder costs families and local services between £11,030 and £59,130 annually.[15] Nearly a third of young people with a severe mental disorder, and 22% of those with a moderate mental disorder, leave full-time education before the age of 15. This compares to 13% for those without a mental disorder.[16] The five percent of children in England with early and severe behavioural problems are 20 times more likely to end up in prison, 6 times more likely to die before 30.[17,18]

1.2.1 Issues with conventional epidemiological study approaches: randomised control trials

Providing evidence to refine existing treatments and identify potentially modifiable risk factors for mental health disorders are complex tasks. Single cause mechanisms in childhood mental disorder are very rare. The "one bug–one drug" model used so effectively in infectious disease treatment and preventative programmes using vaccines, cannot be transposed on mental illnesses.[19] Experimental methods such as randomised trial designs for identifying causal

factors or effective interventions in children are also problematic. In many cases it is impossible to randomise the exposure or intervention under scrutiny – you cannot randomise a child to having autism. Experimental trials require huge resources if they are aiming to detect effects on rare outcomes, such as childhood suicide attempts, or outcomes that occur several years after the intervention.[20] Relative to adults studies, experimental trials in children operate under greater ethical scrutiny and heightened risk concerns from parents, [21] which can lead to rarefied populations being recruited with subsequent limitations on generalisability.[22] The political sensitivities around childhood vulnerability and service inequality can also hinder the applicability of experimental research designs too; a recent example of this involved the UK government being unwilling to adopt scientific recommendations to initially randomise the provision of sure start child care resources. [23]

1.2.2 Issues with conventional epidemiological study approaches: cross-sectional surveys

With limits on the applicability of experimental designs, researchers also rely on well-designed observational [24] or quasi-experiments [25] methods to discern true causes and confounders. From the 1960s, epidemiological approaches have been applied to child psychopathology, and now encompass large-scale cross-sectional surveys and prospective cohorts which integrate biological measures and multi-informant behavioural assessments to unravel aetiological mechanisms.[26] However, these too have their limitations. Nationally representative cross-sectional surveys for child and adolescent mental health disorders are expensive to conduct and occur infrequently – taking the UK as an example, the gap between the previous and current national survey will be 14 years.[7,27] Cross-sectional data also quickly become out of date. Environmental factors, including neighbourhood deprivation, technological advances, schools, and social policy, can change rapidly and swiftly effect patterns of need and detection of child and adolescent psychopathology.[28–30] Furthermore, in isolation these surveys may provide an unreliable picture of the diversity of mental illness prevalence, both through selection bias against mentally unwell populations,[31] and the effects of averaging data, which are representative of the nation, rather than the local or vulnerable populations;[32] children locally resident in Lambeth have very different lives to those in North Norfolk, as do children under the care of the local authority compared to children who reside with their families.

There are also growing concerns that responders to surveys are becoming an increasingly unusual group – response rates are falling across all major surveys especially in younger populations.[33] For example the ONS Labour Force Survey has seen response rates fallen from

73% in 1999 to 43% in 2015, with the youngest age groups (aged 19-25) now significantly underrepresented.[34,35] This may be due to young people leading increasingly busy lives but possibly because of survey fatigue. This is unlikely to improve as more commercial and public sector organisations attempt to gather information from individuals by survey methods.[36] A crucial limitation of cross-sectional surveys, are that they can only be used to describe the relationships between variables at the time of measurement, therefore cannot determine the temporal relationship needed to understand cause and effect. For example, cross-sectional data cannot be used to determine the direction of any causal relationship between childhood depression and obesity.

1.2.3 Issues with conventional epidemiological study approaches: prospective cohorts

Prospective cohort studies have made an extensive contribution to science and the public health, but they too also have several important limitations. They are very expensive to set-up and run. In the United States, a National Institutes of Health (NIH) flagship cohort study, the National Children's Study, started with the remit to garner a nationally representative child cohort over 105 sites. The recruitment and measurement expenses, over the 4-year preliminary phase, amounted to $250,000 per child. By November 2014, the study was stopped due to concerns over the study's feasibility and costs.[37]

The relevance of prospective cohort studies to future generations of young people can be limited. Children born today enter an increasingly digitized world, with greater population movement, economic volatility and greater income disparities.[28,38] The narrow window of childhood development and its susceptibility to changing environmental exposures and demographic shifts, means that new cohort studies need to be commissioned regularly to keep up with socio-cultural contexts of the area they represent. Another issue, particular with paediatric populations, is that once the developmental window has ended, advancement in exposure and outcome measurement [39] and evolved understanding of pre-existing disease risks [40] cannot be retrospectively applied. This makes it harder for research programmes, which use older cohorts to look back at the time of childhood, to produce findings which are applicable to the present.

When childhood mental health outcomes or exposures are relatively rare (<5%) such as exposure to antipsychotic treatment in mid childhood [41,42] or development of an early onset

psychotic disorder,[43] population-based cohort samples are unlikely to include many affected individuals, reducing their power to discern clinically important risk factors and outcomes. Cohort studies can be particularly affected by nonparticipation at given time points, or even complete loss to follow-up. This may be particularly pertinent to psychiatric epidemiology, as nonparticipation is associated with many factors related to child and adolescent mental disorders including non-white ethnicity, socioeconomic adversity, male sex, physical health, cognitive, emotional, and behavioural problems.[44,45] Furthermore, recent work from the ALSPAC birth cohort has shown the genetic predisposition for severe psychiatric disorders (in this case, the polygenic risk score for schizophrenia) is strongly associated with cohort study drop out by age 7.[46] These results suggest that individuals with a genetic predisposition to these schizophrenia and genetically related psychiatric phenotypes, such as other neurodevelopmental disorders [47] will be underrepresented in longitudinal population cohorts analyses. It also implies that longitudinal population cohort studies are likely to be underpowered when examining risks for certain psychiatric disorders. Furthermore, it indicates that some analyses generated from these data may be biased because risk factors and diagnostic outcomes may be not randomly missing.

1.3 WHY DO WE NEED 'BIG DATA' FOR CHILD AND ADOLESCENT MENTAL HEALTH RESEARCH?

In light of the issues described above, conventional survey and cohort based approaches may not provide sufficiently comprehensive approaches to estimate the population need for youth mental health services, the extent to which these needs are being met, and the risk factors contributing to these needs. Again, using England as an example, there are a very limited number of research systems than can provide contemporaneous data on the nature and distribution of child and adolescent mental health (CAMH) disorders in the community. Recent reports by the Children and Young People's Mental Health Coalition and UK Health Select Committee concluded that data used to assess the mental health need of the child and adolescent population across England and the aetiological mechanisms behind these needs, were inadequate.[48,49]

Every school age child living in England, as with many other high-income countries, has a comprehensive digital record, which captures nearly every encounter they have with health, social and education services. These data include individual longitudinal records of birth

details, school performance, physical growth, primary and secondary health care use, psychiatric inpatient use, social and youth justice services contact, employment and training.

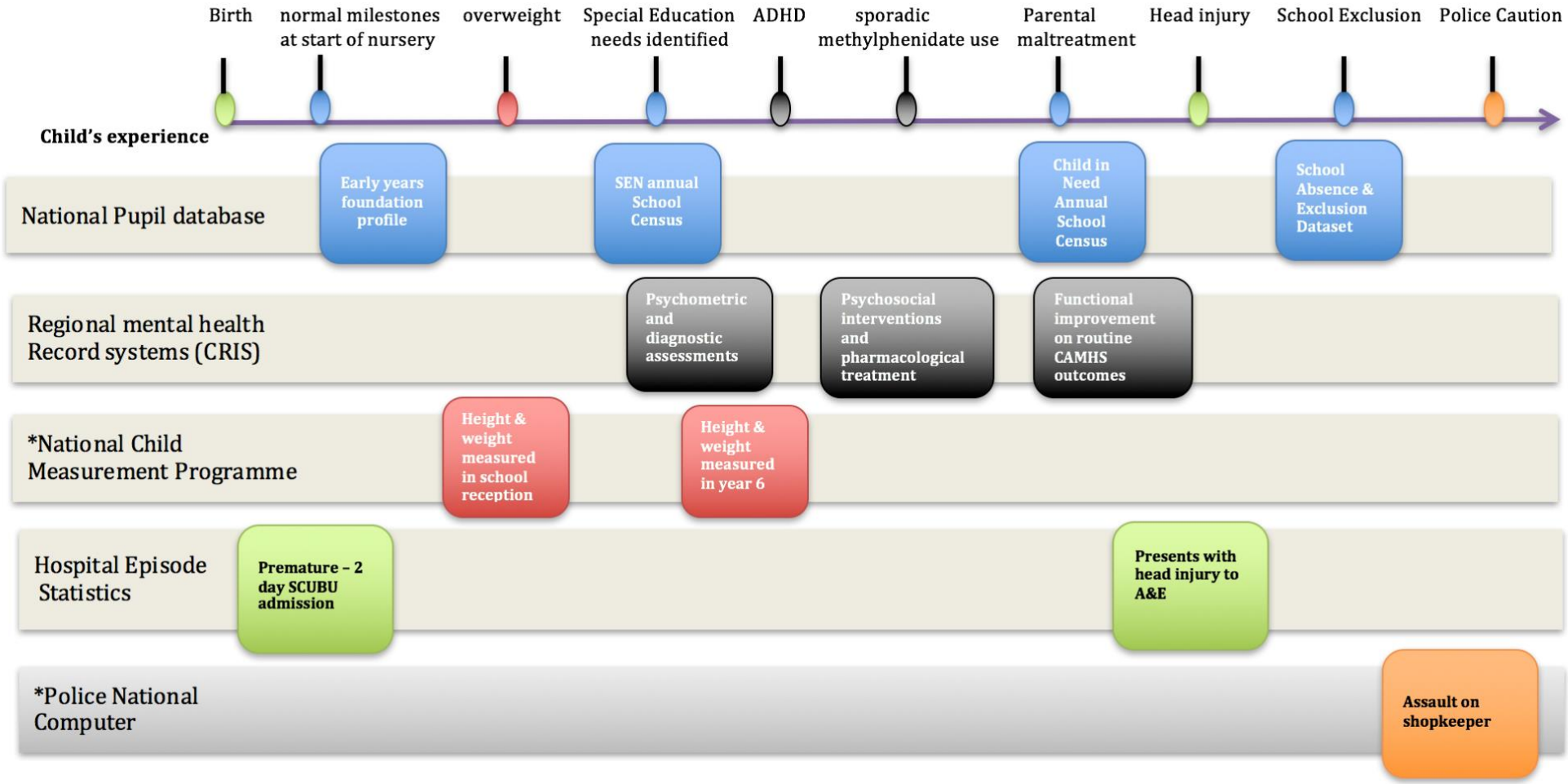
These data are rich, characterising the diverse interactions healthcare staff, social workers, police and teachers have with young people living in their community. In the devolved nations it appears feasible to identify, link and de-identify data to provide an anonymised multi-agency dataset covering youth focused public service activity within existing UK legal and data governance frameworks.[50,51] If replicated across England, or within several large regions, these data could allow researchers to study the risks factors and patterns of disease across very large populations, and provide precise estimates on the outcomes for healthcare interventions. Figure 1.1 provides an example of national and regional databases within England, which have the potential to permit routine analysis of the impact of interventions on key childhood mental health risk factors and outcomes.[52–54] However, these data are complex, representing medical information held in structured, free text, images and video formats, which have not been primarily collected for research. Also, datasets created from the routine administrative outputs of public services are likely to contain more variations in recording and missing information than those completed by small teams working to a clear research protocol. Despite this, such data could afford distinct opportunities that would be difficult to achieve through individually funded research studies – particularly in scale (sample size) and generalisability. Arguably child and adolescent mental health research, an area disproportionately under resourced relative to the individual and societal impact of childhood mental health disorders,[55] may have more to gain from the expansion of these Big Data resources than other areas of healthcare.

1.4 PSYCHIATRIC EPIDEMIOLOGY AND BIG DATA

1.4.1 Data linkage of clinical and social care data

The advent of 'Big data' through the digitisation of mental health and social care information across the world, presents a potentially powerful resource for researchers who wish to study clinical issues "in vivo".[56] However, psychiatric epidemiologists may contend that Big Data approaches are not novel. For decades, Scandinavian countries have led the way in the development of whole population data repositories all linked via a common identification number, acquired at birth or migration to these countries. These repositories can index on an individual level an array of clinical and social information including birth details, school performance, secondary health care use, social and criminal justice involvement.[57]

Figure 1.1 An example timeline of a young person's mental health needs, interventions and outcomes captured in separate large scale longitudinal data sources amenable to data linkage.



*Yet to be linked with routinely collected NHS data

Some countries have developed their use of administrative records to the point where these are now routinely substituted for the more traditional ways of generating data resources for health, social and economic research. In Finland, instead of primary data collection via a census survey delivered to households, thirty different registers and administrative files are linked to provide census data.[58] Outside of Scandinavia, very few other countries are able to do this; the equivalent individual level data exists, but they are contained in separate repositories with no common unique identification number to facilitate linkage between them.

To overcome the lack of a common unique identifier within and across public service systems, two main data linkage methodologies have been developed to create a match for the same individual across separate sets of records.[59] The first is a **deterministic** linkage approach, where a set of predetermined rules are used to classify pairs of records as matched or non-matched. These tend to require an exact or partial agreement on a set of personal identifiers – for example a successful match on the first name **or** surname, **and** match on both the date of birth and postcode. Strict deterministic methods are straightforward to use and commonly employed in government departments, however they can create high levels of missed matches between records.[60] As a consequence, this undermines the confidence that all the relevant records for an individual have been accurately combined across the different data sources.

A second approach to data linkage is **probabilistic** linkage, first proposed by Newcombe in the 1950's,[61] which uses specified identifiers common to each record in both datasets (e.g. surname, post code and date of birth) and, by comparing each set of identifiers in one record against all other records provided, generates a probability estimate of the match being true. So, gender is not particularly informative, as the risk of being matched by chance is 50%, however the chance of 11-digit telephone number being matched by chance on two datasets is very low, hence the probability of a true match on this identifier is high.

To illustrate this with an example: a health researcher needs to link 50,000 patients who have two sets of records A and B held on different data sources. Both A and B hold several personal identifier fields in common such as telephone number, post code, surname and date of birth. An agreement weight can be calculated for each identifier field in record A, depending on the probability (usually an Odds Ratio) of variable matching by chance for the corresponding identifier field in record B. This weight relates to the relative frequency of the identifier in record A and B - in a UK database, the common surname identifier *Brown* will have a lower agreement weight than *Berrycloth*. A total agreement weight can then be calculated for each A

and B pairing across all the identifiers for each patient. These AB pairings are then ranked in probability and a cut off determined, often through manual review, to select the lowest probability score taken to represent a positive match for a record pair. This process is computationally intensive: for 50,000 patients there are 1.25 billion unique A + B record pairings. Hence, efficient probabilistic linkages across very large health and social care datasets have only become feasible in the last decade.[54]

1.4.2 Natural language processing and the Electronic Health Records

Linking data across very large scale administrative databases can provide a very powerful resource for epidemiological studies. However, research designs based on these linked databases have limitations which are important to consider. As described in the first paragraph of this chapter, a major strength of these administrative data is their comprehensive inclusion of the whole population of interest, and therefore providing highly generalisable results. This is counterbalanced, however, by the 'Veracity' or quality of these data, which can be problematic.[57] This was illustrated by a recent critique of the UK Department of Health's Hospital Episode Statistics (HES), the vast database which contains details of every NHS hospital inpatient admission, emergency department and outpatient contact in England. In an analysis of HES data, the authors revealed some impossible events. In a one year period (2009-2010) there were over 17 000 male inpatient admissions to obstetric services, over 8000 to gynaecology outpatients with nearly 20 000 midwife episodes.[62] Admittedly the proportion of these errors were small relative to the scale of the HES record system, which adds 125 million admitted patient, outpatient and accident and emergency records each year to its database.

Concerns around the veracity of administrative data reduces confidence in the validity of the findings generated from the data. There are ways of managing this. Researchers using administrative data have determined the extent of misclassification and whether it is systematic, by undertaking a validation exercise with a subset of patients. This may involve cross checking the same clinical variables using linked patient level data in HES and other independently administered case registers.[63] Although these 'impossible scenarios' or internal errors are readily detectable within the database itself, some key clinical data (e.g. presenting complaints, diagnoses, treatments received) cannot be cross-checked within the administrative database alone to test their validity.[64] This limitation also reveals that researchers only have access to variables which have operational value (i.e. pertinent to clinical or administrative practice) rather than those which are most relevant to the risk factor or health outcome under

investigation, and highlights the potential for a superficial characterisation of individuals within administrative health databases.

Some authors conclude the advantage of very high sample populations in administrative mental health, will always be counteracted by low data validity and reliability.[57] But, as Robert Stewart in a recent editorial argues, this doesn't have to be the case. He describes an approach which goes behind the structured data held within administrative health to the source health data itself, and allows exploration of the unstructured text within electronic mental health records.[65] As he describes, this method can bring the benefits of statistical power and generalisability of large scale administrative data but also reduce the limitations of the low data validity and data reliability; essentially, switching the coding processes of administrative health data from expert coders back to clinicians.[66]

Clinical notes reviews have a considerable history in mental health research, from the studies of asylum records in the Victorian era [67] through to the growth of the 'case register' in the mid- to late twentieth century.[68] Clinical notes often provide a detailed account of the patient's symptoms, treatments and outcomes, avoiding the recall bias and patient burden often associated with information gathering from standard primary data collection survey approaches. However, clinical notes within psychiatry are often extensive, and it takes time for those with suitable expertise to manually screen, annotate and extract research data. As such, research using clinical note reviews have remained limited by small sample sizes and the potential for high inter-observer variability.[69]

In the last 2 decades, a number of technical advances have facilitated the use of clinical notes for research. The advent of computerised or electronic health records means that clinical notes were able to be transferred and manipulated by digital systems.[70] This process was initially welcomed by health workers as method to finally operationalise and structure clinical notes. There was an expectation that electronic records would provide large clinical samples to exploit for use in research.[70] In mental health, research systems were adapted by researchers, such as the OPCRIT system which aimed to improve the objectivity of diagnosis in NHS clinical settings, and provide a mechanism for the routine collection of a core clinical, research and audit data.[71,72]

Despite electronic records becoming nearly ubiquitous across NHS mental health services[73], their potential remains to be fully realised. Public concern over the ethical use of secondary

use of health care records [74] and the complexities of the records themselves means they are currently not delivering their potential for research. The data complexities remain, as despite the considerable effort to operationalise structured note keeping into mental health records,[72] it has failed to be adopted into routine practice. Clinicians tend to shun structured templates or drop-down options for the majority when keeping a record of the daily practice. An overview on why the free-text note persists as the predominant method of recording clinical information suggests:[75] free text is viewed by clinicians as a convenient method of expressing clinical concepts and events, such as diagnosis, symptoms, and interventions;[76] prose can be more accurate, reliable, and understandable;[77] free text is tolerant of ambiguity, which supports the complexity of clinical practice;[78] medical notes are nuanced and makes heavy use of negation (e.g. "she denied any current suicidal ideation") temporal expressions ("symptoms resolved a few months ago"), and hedging phrases ("the treatment was somewhat successful"). All of these important elements are difficult to represent as categorical options within a structured form.

1.4.3 Applying Natural Language Processing to Electronic Health Records

Computational linguistics or Natural Language Processing (NLP) is an area of research and application which explores how to make computer systems understand and manipulate natural language expressed in text to perform desired tasks.[79] NLP techniques have now evolved sufficiently to rapidly process and interpret the wealth of contextual, unstructured health data held within electronic records.[80] Two powerful applications of NLP have been deployed in health research: one includes de-identifying electronic health records, essentially scrubbing personal identifiers within the free-text of a patient record;[81] another has been its use in information extraction.

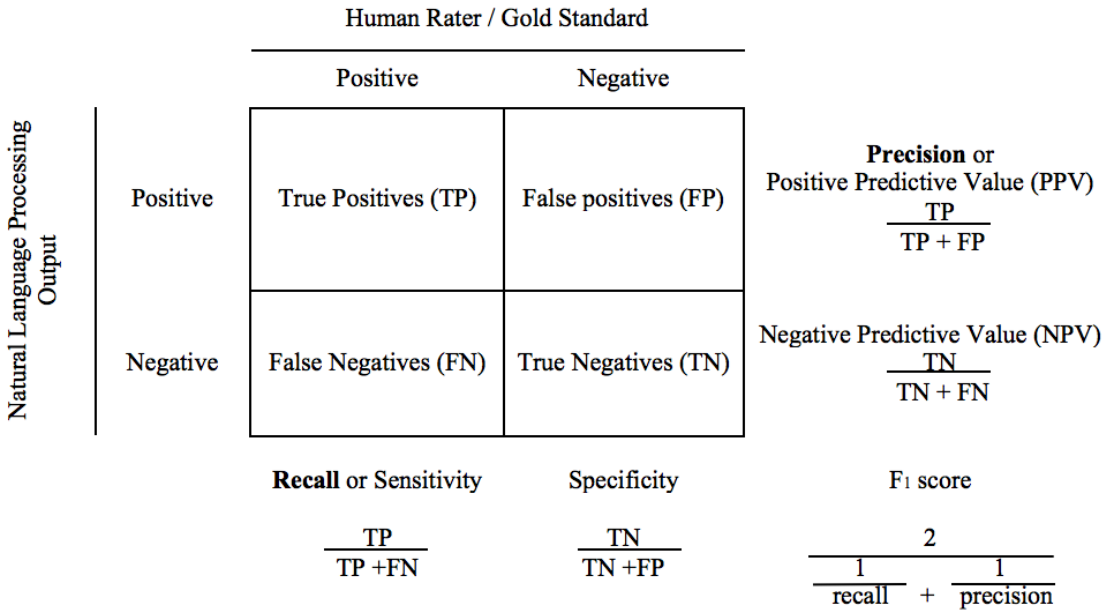
NLP can be used to discern the meaning or semantic content of text, and using pre-specified algorithms, can encode this text to provide structured output for analysis. This provides considerable advantages compared to performing key word searches in health data. For example, key word searches on the term *suicide* will provide every mention of the term in the health record and not discriminate whether it references a patient's history, describes a current mental state or was just part of clinician screening. This approach is difficult to use for any large scale analysis, as they require manual review to add context.

Two different NLP approaches are generally used in health data extraction, rule-based and machine learning, sometimes separately or in combination to identify a desired phenotype or event in clinical text. Rules-based NLP relies on human or expert consensus to arrive at a protocol of how a combination of text based terms (clinical or otherwise) may be combined with logic rules (via AND, OR, and NOT) in order for a particular phenotype or event to be positively or negatively identified. This set of rules are then translated into an NLP algorithm and used to detect cases, exposures or outcomes of interest over the health record. This type of approach is currently being implemented by eMERGE, a very large US national network, across 9 regions, with two clinical sites specifically for paediatric populations which combines DNA biorepositories with electronic medical record (EMR) systems for large scale, high-throughput genetic research.[82] Machine learning NLP approaches use pattern recognition via statistical or machine learning methods to identify a phenotype or exposure of interest within the free text records. Confidence parameters around accuracy can be stipulated, allowing uncertainty on whether an event or phenotype is a true positive, which can be accounted in later analysis. Investigators have largely adopted this approach in i2b2 (Informatics for Integrating Biology and the Bedside), a US consortium, based at Harvard/MIT Health Science division and Partners HealthCare System in Boston, Massachusetts.[83]

Testing whether a NLP tool accurately identifies clinical text for later analysis, is very similar to evaluating the accuracy of a new diagnostic or screening procedure. The NLP tool output is compared to a gold standard output (see figure 1.2) often created by a manual review of the same text. NLP accuracy is measured in terms of its precision (positive predictive value) and recall (sensitivity). Precision is simply the terminology used within the NLP field for Positive Predictive Value (PPV), which refers to the proportion of true positive terms or sentences classified by the NLP application out of the total number classified as positive by the NLP application. Recall (a NLP term which can used interchangeably with Sensitivity) refers to the proportion of true positive sentences classified by the NLP application out of the total number of true positives. Thus, an application with a high degree of precision is necessary to reduce the frequency of false positive classifications and a high degree of recall is necessary to reduce the frequency of false negative classifications. As with screening tests there is often a trade-off between recall and precision, and clinical researchers often need to make a decision whether their requirements should lean towards having high precision at the cost of not capturing all diagnostic instances, or high recall where all potentially relevant terms will be captured but many more will false positives. A commonly used measure within NLP which conveys the

NLP overall performance or accuracy is the F1 score, which is calculated as a weighted average of the precision and recall, with the most accurate value at 1 and lowest at 0.

Figure 1.2: Framework for evaluating the accuracy of a NLP application to clinical notes



In adult samples, NLP has been deployed within the electronic health records, to support a number of clinical epidemiological studies, including the examination of factors associated with adverse drug responses,[84] surgical complications [85] and treatment resistant depression.[86] Both rule-based and machine learning approaches require clinical raters to produce gold-standard data sets, which depending on the complexity of the task, can take considerable resources to produce. The main advantage of the statistical or machine learning approach over rule-based approaches are that identification/ categorisation of the target features (i.e. the event or phenotype of interest within the notes) are purely data driven, with no expert consensus required on how rules are constructed. The main disadvantage is the lack of transparency of the data driven process - machine learning does not provide an accessible flow of logic procedures, which makes it difficult to trouble shoot when accuracy is below satisfactory, or it requires adaptations for use in other databases. Figure 1.3 below provides a hypothetical example of how an NLP approach may be applied to a clinical epidemiological study.

Figure 1.3 A hypothetical example of how NLP may be applied to epidemiological research

A research team wished to establish the number of primary and secondary psychiatric diagnoses for patients after admission to a psychiatric inpatient unit. They had a list of 40,000 patients with electronic records which covered 10 years post hospital discharge. They decided it was not practical for these to be manually reviewed on a large scale, so they employed a rule-based approach to build a NLP application. A coding framework was devised to capture unique diagnostic psychiatric categories, with variations of classification subtypes, abbreviations spellings and misspellings accorded to each category. This framework was created from a consensus of experts who were familiar with the recording of diagnostic information in the health record. Clinical raters then applied this framework to batch of clinical notes which were randomly extracted from a random subsample of study patients. They used NLP software to annotate and categorise diagnostic terms from the health record, which manually imposed structured data on the free text. This produced a 'gold standard' of manually ascertained diagnostic annotations. The gold standard batch was then split into a 'training' set and a 'test' set. Meanwhile, researchers using NLP software created a set of rules/algorithms which automated the clinical coding framework, to take advantage of previously published diagnostic classifications and clinical ontologies.

After the first test where the NLP tool diagnosis output was compared against the gold standard training set, the NLP performance was application underwent a number of iterations. The clinicians previously decided that precision and recall above an arbitrarily set value of 0.8 would suffice So, the clinical reviewers examined the diagnostic classification errors by examining False positive or False Negative sentences or documents. They noticed particular patterns in the misclassification – diagnostic terms omitted from the ontology, terms used to negate the presence of a diagnosis - which they corrected by making changes to the ontology, and creating some additional linguistic rules to reclassify terms within the text. They then ran this new iteration against the training set and evaluated its performance. This process was repeated until the performance metrics were satisfactory. The NLP tool was then compared against the gold standard test set and performed well, and so extended to extract diagnostic information from the whole sample.

The research team were also interested to see whether a NLP machine learning approach performed as well as the rules based NLP tool to detect the diagnostic construct of "psychosis". They applied a statistical approach called a support vector machine (SVM) to 'learn' the position of "psychosis" related key words in the gold standard training set of documents. The SVM approach used all the words, spaces, punctuations within each training document as multidimensional data points to build statistical models which predicted the probability of a "psychosis" key word being a clinical confirmation of a psychosis diagnoses. After the learning phase was complete, the SVM models derived from the training data, were applied to the test data. The research team were then able to compare the performance metrics of the rule based and machine learning NLP applications. The eventual outputs from both NLP applications, were run over a large corpus of patient notes, to produce tractable, time stamped diagnostic information for each document, with all text occurrences of diagnoses in the text annotated by the NLP software.

1.4.4 Natural language processing and its application in child and adolescent psychiatric epidemiology

To understand how natural language applications have been applied to big data resources in child and adolescent mental health research, I conducted a literature search of the MEDLINE, EMBASE and Psychinfo databases using OVID Gateway.[87]

As outlined in figure 1.4, the search strategy used a recent comprehensive review to build a list of articles published over the last 10 years (up to 20 June 2017) from 85 'big data' health resources.[56] Using filtering terms, I restricted these 'Big Data' articles to those using NLP applications to investigate risk and outcomes relevant to child and adolescent populations with mental health disorders. NLP search terms were identified from a recent systematic review on text mining.[88]

Initial searches revealed a number of studies were using combined health administrative data, for example health insurance and prescribing registers, or structured fields within primary care records to conduct psychopharmaco-epidemiological studies within child and adolescent populations. However, after imposing the natural language search terms and excluding articles generated from this thesis (see figure 1.5) 329 published articles remained. After an abstract and title review, and a full text article assessment, only 9 articles, from 4 research groups in North America (tables 1.1 and 1.2) and one from France (table 1.3), revealed any studies published within the last 10 years related to child and adolescent mental health, and using NLP applications within the Big Data resources described in figure 1.4.

Figure 1.4 Search terms used to identify studies of NLP applications within Big Data resources in child and adolescent mental health

Big Database resources search terms [56]

Israel's psychiatric case register or Clalit Health Services or Hong Kong Hospital Authority or (Seoul National University and Mental health) or Taiwan National Health Insurance Database or (Mental Health National Outcomes and Casemix) or Western Australian data linkage system or Asturias Cumulative Psychiatric Case Register or Gmunder ErsatzKasse or German Research Network on Depression or (Health Search Database and Italy) or French National Health Insurance Fund or DGPPN-BADO or South Verona Community-Based Mental Health Service or (Zurich and Psychiatric case register) or ((Clinical Practice Research Data link or CPRD or GPRD) and General Practice Research Database) or Clinical Record Interactive Search or Generation Scotland or (Galatean risk and safety) or mental health minimum dataset or QResearch or The Health Improvement Network or UK Biobank or Secure Anonymised Information Linkage or PsyCymru or Danish Psychiatric Central Research Register or deCODE Iceland or Dutch National Survey in General Practice or Finnish Hospital Discharge Register or (Netherlands and Psychiatric Care Register) or Norwegian Patient Register or Odense University Pharmaco-epidemiologic Database or (Odense University and Database and pharmacology) or Hungarian National Health Insurance Fund or European Observatory on Health Systems or European Autism Interventions or (prescription database and Norway) or PROTECT-EU or eDESDE-LTC or Canadian Chronic Disease Surveillance System or Canadian Primary Care Sentinel Surveillance Network or OntarioMD or Ontario Mental Health Reporting System or Saskatchewan Health Databases or 23andMe or (Healthcare Cost and Utilisation Project) or Data QUEST or (Electronic medical records and genomics network) or Group Health Research Institute or (Health Plan Employer Data and Information Set) or (Informatics for integrating biology and the bedside) or (Centers for Disease Control and Prevention) or (KP Research Program on Genes, Environment and Health) or (Kaiser Permanente and mental health) or (Mayo Clinic and mental health) or MarketScan Research Database or (mental health and medicare) or Health Care Systems Research Network or National Prescription Audit or (National Disease and Therapeutic Index) or New York Presbyterian or Palo Alto Medical Foundation or SHRINE or Scalable Partnering Network for Comparative Effectiveness Research or Stanford Translational Research Integrated Database Environment or Texas Department of Criminal Justice or University of Michigan Health System or Vanderbilt University Biorepository or Veterans Affairs Database or Asian Pharmacoepidemiology Network or Psychiatric Genomic Consortium or IMS Prescribing Insights database or WHO Global Health Observatory Data Repository)

Natural Language Processing search terms [88]

text mining or literature mining or machine learning or machine-learning or automation or semi-automation or semi-automated or automated or automating or text classification or text classifier or text categorization or text categorizer or classify text or category text or support vector machine or SVM or Natural Language Processing or active learning or text clusters or text clustering or clustering tool or text analysis or textual analysis or data mining or term recognition or word frequency analysis

Child and adolescent search terms

children or child or paediatric or adolescence or adolescent

Mental Health search terms

mental health or psychiatry

Figure 1.5 Flowchart of study inclusion criteria

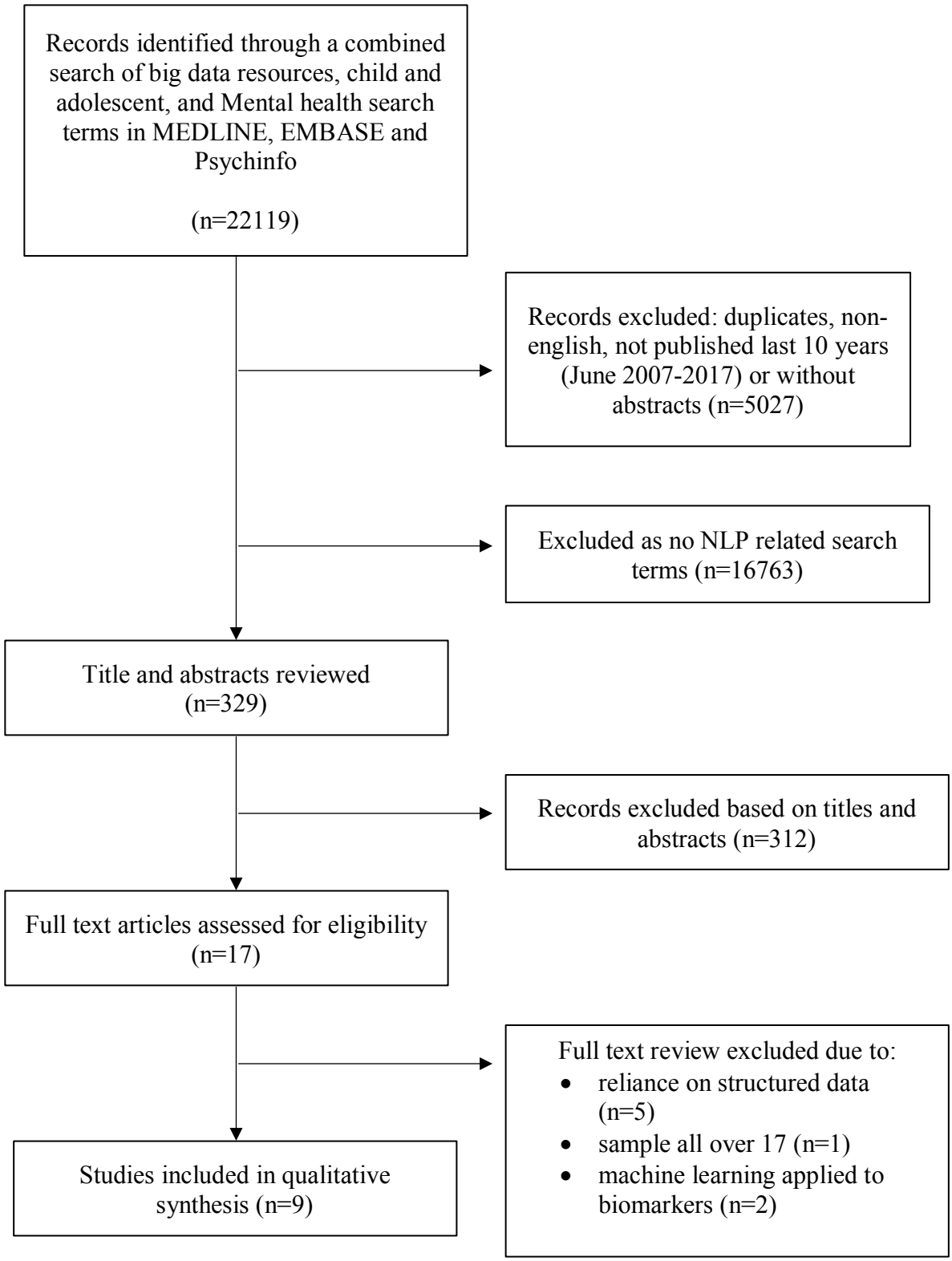


Table 1.1 Summary of included studies using Big Data resources and NLP applications for child and adolescent mental health research: Harvard

First author, year,	Institution	Aim and design of study	Data Source Types	Sample	Age range	NLP approaches[a]	Results
Clements et al. 2015 [89]	Harvard/MIT Health Science division and Partners HealthCare System in Boston, Massachusetts	Nested case-control within a retrospective cohort linking maternal records via matching child's date of birth and surname, insurance identifiers, and hospital encounter date. Examine risk of neurodevelopmental associated with perinatal exposure of antidepressants	ICD-9 coded related to ASD and ADHD, RxNorm Prescription codes, disorders derived from structured and free text within the clinical notes	1,377 children with ASD matched to 4,022 healthy control children and 2,243 with ADHD (but no ASD diagnosis) matched to 5,631	children age 2–19	**Rule Based NLP Algorithm:** example terminologies include International Classification of Diseases (ICD), National Drug Code (NDC), and Logical Observation Identifiers Names and Codes (LOINC).	ASD, sensitivity is 1.00, specificity 0.91; for ADHD, sensitivity is 0.84, specificity 0.90) antidepressant exposure prior to and during pregnancy was associated with ASD risk, but risk associated with exposure during pregnancy was no longer significant after controlling for maternal major depression [O.R 1.10 (0.70–1.70)]. Antidepressant exposure during associated with ADHD risk, even after adjustment for maternal depression
Castro et al., 2016 [90]	Harvard/MIT Health Science division and Partners HealthCare System in Boston, Massachusetts	Replication study: Nested case-control within a retrospective cohort linking maternal records via matching child's date of birth and surname, insurance identifiers, and hospital encounter date Examine risk of neurodevelopmental associated with perinatal exposure of antidepressants	ICD-9 coded related to ASD and ADHD, RxNorm Prescription codes, disorders derived from structured and free text within the clinical notes	1245 ASD cases matched to 3,735 healthy controls and 1701 ADHD cases matched to 5103	children age 2–19	**Rule Based NLP Algorithm:** example terminologies include International Classification of Diseases (ICD), National Drug Code (NDC), and Logical Observation Identifiers Names and Codes (LOINC).	No significant increased risk for ASD and ADHD associated with prenatal antidepressant exposure. Risk associated with pre-pregnancy antidepressant exposure, and with prenatal maternal psychotherapy. Supporting confounding by indication as possibility.

[a] References the mapping exercises where text is categorised according to standardized nomenclatures - clinical terms (e.g. UMLS, ICD-9 SNOMED-CT) medications (e.g. RxNORM, National Drug Code) laboratory observations (e.g. Logical Observation Identifiers Names and Codes)

First author, year,	Institution	Aim and design of study	Data Source Types	Sample	Age range	NLP approaches [a]	Results
Doshi-Velez et al., 2014 [91]	Harvard/MIT Health Science division and Partners HealthCare System in Boston, Massachusetts	Cross sectional analysis of retrospective records. Patterns of co-occurrence of medical comorbidities in ASDs	ICD-9 coded related to ASD and other clinical disorders derived from structured and free text within the clinical notes	4934 individuals with ASD, a tertiary-care paediatric hospital (mean follow-up 11 years, SD 4.8 years),	Samples at least over 15, paediatric history examined	**Rule Based Algorithm:** example terminologies include International Classification of Diseases (ICD), National Drug Code (NDC), and Logical Observation Identifiers Names and Codes (LOINC).	Clustering analyses revealed 3 high-morbidity subgroups: 1 characterized by seizures, 1 characterized by psychiatric disorders, and 1 characterized by more complex multisystem disorders.
Kohane et al., 2012 [92]	Harvard/MIT Health Science division and Partners HealthCare System in Boston, Massachusetts	A retrospective prevalence study across three general hospitals and one paediatric hospital. Examine patterns of co-occurrence of medical comorbidities in ASDs	ICD-9 coded related to ASD and other clinical disorders derived from structured and free text within the clinical notes	14,381 individuals with ASD	under age 35	**Rule Based NLP Algorithm:** example terminologies include International Classification of Diseases (ICD), National Drug Code (NDC), and Logical Observation Identifiers Names and Codes (LOINC).	Burden of co-morbidity is substantial and present across multiple health care systems with over 10 percent of patients with ASD having bowel disorders, or epilepsy, over 5% with CNS or cranial anomalies, and over 2% with schizophrenia.

Table 1.2 Summary of included studies using Big Data resources and NLP applications for child and adolescent mental health research: Other US

First author, year,	Institution	Aim and design of study	Data Source Types	Sample	Age range	NLP approaches [a]	Results
Lyalina et al., 2013 [93]	Center for Biomedical Informatics, Stanford University, USA	Cross sectional analysis of retrospective records. Elucidate the phenotypic boundaries of autism, bipolar disorder, and schizophrenia. Examine individual-level phenotypic variation within each disorder, as well as the degree of overlap among disorders.	Hospital/Community EHR systems, incl. ICD-9 codes, RxNorm Prescription codes, Procedure codes, pathology reports, radiology reports, and transcription reports.	7000 patients at two facilities with ASD, Schizophrenia or Bipolar diagnoses	Aged 15 +	**Rule Based NLP Algorithm:** Clinical ontologies used to build Concept Unique Identifiers (CUIs) from the Unified Medical Language System (UMLS) Metathesaurus, ICD-9 codes, treatment codes; additional filters remove ambiguous terms, flag negated terms and terms attributed to family history. LASSO Logistic regression models trained to predict diagnosis of depression, response to treatment and severity.	Principal component analysis isolated autism as a separate disorder, while revealing significant overlap between schizophrenia and bipolar disorder.
Huang et al., 2014 [82]	Center for Biomedical Informatics, Stanford University, USA	Retrospective cohort design. Develop and evaluate computational models that use electronic health record (EHR) data for predicting the diagnosis and severity of depression, and response to treatment.	Hospital/Community EHR systems, incl. ICD-9 codes, Prescription codes, Procedure codes, pathology reports, radiology reports, and transcription reports.	35 000 patients (5000 depressed) from the Palo Alto Medical Foundation and 5651 patients treated for depression from the Group Health Research Institute	not specified, likely to include children	**Rule Based NLP Algorithm:** Clinical ontologies used to build Concept Unique Identifiers (CUIs) from the Unified Medical Language System (UMLS) Metathesaurus, ICD-9 codes, RxNORM, treatment codes; additional filters remove ambiguous terms, flag negated terms and terms attributed to family history.	Area under curve for diagnosis: 0.8 (95% CI 0.784-0.815) at 90% specificity, sensitivity is 50% at time of diagnosis.

Anderson et al.,2015 [94]	Department of Clinical Pharmacy, University of Colorado, USA	Retrospective analyses of de-identified EHR data of primary care organizations to estimate the frequency of using diagnostic codes to record suicidal ideation and attempts.	ICD-9 Codes, clinician notes, suicidal ideation items; and diagnostic codes from the EHR.	61,464 patients with a new episode of depression	Aged 15 +	**Rule Based NLP Algorithm:** ICD-9 coded diagnoses, free text positive mention or negation of suicidal ideation in history presenting illness (HPI) fields, and PHQ-9)	HPI data (n = 15,761), 1,025 had a free-text indication of suicidal ideation recorded in their HPI. 3% (n = 30) had a corresponding ICD-9 code indicating suicidal ideation
Lingren et al., 2016 [82]	Cincinnati Children's Hospital Medical Center, Division of Biomedical Informatics, Cincinnati, Ohio, USA: eMERGE Network	Cross sectional analysis of retrospective records. Developing an automated algorithm and comparing rule based vs machine learning applications for extracting cohorts, and examining the co-occurrence patterns of comorbidities associated with patients with ASD	ICD-9 codes and ASD concepts derived from the free text within clinical notes	14,758 and 4,229 patients with ASD, from the Boston Children's Hospital (BCH) and Cincinnati Children's Hospital Medical Center (CCHMC) EHR databases respectively	Age not specified	**Rule Based NLP Algorithm:** UMLS using the Apache cTAKES natural language processing system. The default cTAKES dictionary (UMLS SNOMED-CT and RxNORM pruned by semantic types for Diseases/Disorders, Signs/Symptoms, Anatomical Sites, Medications and Procedures) enriched with the ASD terms. **Machine learning** (Support vector machine) using the EHR inputs also applied.	The rule-based better on BCH data (BCH, 0.885 PPV; CCHMC, 0.840 PPV), Machine learning algorithm performed similarly at both sites (BCH, 0.780 PPV; CCHMC, 0.799 PPV). 3 high-morbidity subgroups: 1) psychiatric problems 2) developmental disorders including dyslexia, lack of coordination, and various disorders of the ear, skin and other bodily systems; 3) epilepsy and recurrent seizure.

[a] References the mapping exercises where text is categorised according to standardized nomenclatures - clinical terms (e.g. UMLS, ICD-9 SNOMED-CT) medications (e.g. RxNORM, National Drug Code) laboratory observations (e.g. Logical Observation Identifiers Names and Codes)

Table 1.3 Summary of included studies using Big Data resources and NLP applications for child and adolescent mental health research : Rest of the world

First author, year,	Institution	Aim and design of study	Data Source Types	Sample	Age range	NLP approaches [a]	Results
Metzger et al., 2017[95]	Paris-Sorbonne University, Hôpital de la Croix-Rousse, Lyon	Nested case control study in one hospital EHR. To compare NLP methods and the national surveillance system for suicide attempt rate detection in France with Emergency Department records	ICD-10 (structured) principal and associated diagnoses free text EHR notes from the ED	307 cases: 614 controls	Aged 15 +	**Machine learning approach**: UrgIndex, servlet identifies free-text medical terms to a French-language medical multi-terminology indexer (ECMT), additional filters remove ambiguous terms, flag negated terms. All medical terms are then mapped onto the UMLS meta-thesaurus. Range of Machine learning (SVM, neural network, decision trees) approaches used to identify suicide related EHR inputs	Gold standard suicide prevalence 4.8%. National surveillance network 0.74% Depending on machine learning approach prevalence was 4.6 to 11.4% (F-measures 70.4 to 95.3) according to machine learning approach. Best NLP approach significantly more accurate than national surveillance rates

Almost all these studies used NLP to discern variations in ASD phenotypes and their respective clinical correlates from electronic health records. For example, both groups at Harvard Medical School and Stanford University applied NLP extraction processes across local secondary care health record systems. The Harvard group examined the physical and mental health comorbidities of 14,000 children and adults with ASD over a 15-year period. From this work they suggested that ASD associated co-morbid symptoms and disorders could all be clustered as related to neuronal and synaptic function, including increased seizure frequency, sleep disorders, bowel disorders and schizophrenia.[91,92] Stanford University used NLP approaches in 7000 adult and child patients to identify discrete and overlapping phenotypic 'signatures' for ASD, Schizophrenia and Bipolar disorders within the health record. They concluded that this technique should enable researchers to build research cohorts of patients who meet similar phenotypic criteria for a particular disease.[93] These studies were looking to produce novel perspectives on how physical and mental disorders cluster. However, many were limited by only examining diagnostic terms, rather than symptoms or the severity of the particular conditions. Furthermore, the accuracy of the extraction methods used (i.e. performance metrics against a gold standard) was not always available. In addition, service level variation in data quality was not examined, therefore spurious associations between conditions could have occurred because one particularly service in a hospital was good at documenting comorbidity relative to another. The only non-US study using NLP in large scale clinical data, examined the accuracy of a machine learning approach to identify presentations of suicide attempts to an Emergency Department at a major hospital in Lyon, France. The authors found Bayesian approaches provided the greatest accuracy with prevalence rates of 4.9%, very similar to the manually rated gold standard detection, which estimated the suicide rate at 4.8%. They concluded that NLP approaches could supplement current national surveillance methods (this approach provided a prevalence rate of 0.8%) which provided an estimate 6 times lower that the true prevalence. However, these findings were derived from a single site, using French vocabulary in a small sample, potentially limiting its generalisability. This NLP approach will require replication and testing in other hospitals before it can be more widely adopted for national surveillance, and will need significant adaptation if applied to non-French language healthcare systems.

1.4.5 Examples of other Big data methodologies

In respect to other Big Data methodologies beyond data-linkage and Natural Language Processing, a recent narrative review revealed a number of novel approaches are being tested to see whether they can provide novel, time efficient methods of measuring psychiatric symptom prevalence, trends and treatment outcomes.[96] To give a few examples: The Durkheim Project is attempting to measure the prevalence of suicidal feelings in army veterans from text mining within clinical notes and social media.[97] Online resources, such as PatientsLikeMe [98] and myhealthlocker [99] are assessing the safety and effectiveness of treatments related to a number of psychiatric disorders using patient-reported outcome data. Mobile phones are being used for continuous monitoring, via their internal accelerometers, to detect abnormal movement patterns in patients with Attention Deficit Hyperactivity Disorder [100] and Parkinson's disease.[101] Big Data technologies are also being used to conduct 'agnostic' analyses, where statistical algorithms process huge volumes of data in order to detect previously unrecognised relationships or hidden signals between exposures and disease outcomes. In this process, analyses are looking to generate new hypothesis or predictive models, rather than test existing theories of causal relations between exposure and outcome variables. These 'hypothesis free' analyses are now an integral part of the post-marketing safety evaluation performed by pharmaceutical regulators, who scan huge volumes of surveillance data to detect correlations between drugs and adverse events.[102] The hope for these projects, and many similar, is that Big Data methodologies may eventually be integrated into clinical records to enhance the evaluation of clinical interventions and service provision.

1.5 AIMS AND STRUCTURE OF THIS THESIS.

As described in the literature review, there has been a very limited number application of big data techniques which combine large scale data linkage and NLP approaches to clinical epidemiological studies of child mental health. As it is likely, at least within the foreseeable future, that administrative data resources will continue to grow, and the reliance of free text in electronic records will continue as the principle tool to communicate across youth orientated services, there is a need to examine how health care records can be exploited by novel big data techniques including data linkage and NLP methods to extend conventional epidemiological approaches in children and adolescents mental health. Furthermore, there is a need to understand what are the methodological and governance processes that may facilitate or

impinge on such use. Encompassed within this theme, the thesis contains a collection of self-contained chapters, which all aim to test distinct hypotheses, summarised as follows:

Chapter 2

Children with autism spectrum disorders (ASD) are more likely to receive antipsychotic medication than any other psycho-pharmacological treatment. Prior work has established that a number of co-morbid conditions were associated with antipsychotic use in ASD, but a number of methodological issues limit the conclusions regarding which symptoms clinicians target when they prescribe antipsychotics to ASD children. In this study, I used electronic health records and NLP data extraction techniques to determine variables of interest to better exploit the information held in both structured and unstructured text. I undertook a retrospective cohort study to examine which psychiatric co-morbidities were associated with antipsychotic use, after adjusting for a number of other potential confounders including aggression and self-injury, which are the main symptomatic indicators for antipsychotic treatment.[103]

Chapter 3

During the analysis for study described in chapter 2, I became aware that many of the structured risk assessments within the clinical record were inadequate in distinguishing risk of suicide/ self-harming behaviours from the self-injurious behaviour associated with functional impairment and stereotypical behaviour in ASD.[104] This prompted the study in this chapter, to investigate whether a recently constructed NLP tool for detecting suicidal related references in the free-text, could be adapted and validated within the ASD sample described in chapter 2.

Chapter 4

This chapter builds on the methodology in chapter 2, but this time describes a study examining the effect of ASD as a co-morbidity on anti-psychotic treatment profiles in young people (i.e. age under 18) with early onset psychosis. In this longitudinal study, I aimed to investigate whether co-morbid ASD was associated with a pragmatic measure of poor antipsychotic treatment response in a large historical clinical cohort of children and adolescents with first-episode psychosis. Previous studies had demonstrated that pre-morbid adjustment disorders within clinical samples were associated with poor prognostic factors in early onset psychosis. This suggested that the effect of specific neurodevelopmental conditions, such as ASD which, by definition, represents extreme manifestations of poor premorbid difficulties, may be associated with poor response to antipsychotics. In this study, I tested the hypothesis that

young people with co-morbid ASD would be more likely to experience antipsychotic treatment failure.

Chapter 5

This chapter describes how I adapted and tested a recently validated NLP approach to extract negative symptoms of psychosis from the free text records of the early-onset psychosis cohort first introduced in chapter 4. Negative symptoms (NS) are an important prognostic risk factor established in adult-onset samples but rarely examined in early onset samples. To explore NS as potential prognostic indicator early onset psychosis, I examined whether NS at first episode predicted antipsychotic treatment failure. Work in adult-onset samples, suggests that NS characterizes psychotic disorders with non-dopaminergic pathophysiology, and lower responsiveness to current antipsychotics which block Dopamine receptors. Therefore, I predicted that children and adolescents with NS at presentation would be more likely to experience antipsychotic treatment failure. I also expected that this association would remain after taking account of potential confounders, including type of psychotic disorder, co-morbid depression, and additional markers of premorbid neurodevelopmental difficulties such as co-occurring autism spectrum disorders, hyperkinetic disorder and intellectual disability.

Chapter 6

One of the main limitations of only using secondary healthcare clinical health records to conduct epidemiological research, is the lack of a control population to reference (i.e. non-cases). Without accurate population denominators, it is difficult to estimate diagnostic prevalence and incidence rates. Also, without a control group, it is difficult to estimate the effect of population based risk factors for psychiatric disorders, and the effect of psychiatric disorders on non-health related outcomes. In this chapter I provide an overview of the work I undertook to overcome this by establishing the first linkage in England of routinely collected data between a large local regions child and adolescent health, education and social care services. In this chapter, I describe how the Clinical Record Interactive Search (CRIS) programme was used to join up data from health, education and social services for children living in four local authorities in South London to create two datasets: one linking acute general hospital data (NHS Digital Hospital Episode Statistics, HES) to children's mental health services and the second linking mental health data to education data (Department for Education National Pupil Database). I describe these resources, give examples of how they can be used

to improve services, and discuss what is needed to implement this approach more widely across the UK.

Chapter 7

Using the CRIS linkage to the National Pupil Database (NPD) as a case example, this chapter provides a comprehensive account of the ethical, legal and technical challenges of accurately linking routinely collected public service data to examine associations between childhood mental health disorders and school performance. I provide more detail on the samples of both CRIS and NPD datasets, the governance paths undertaken to establish the ethical and legal framework for establishing the linkage, the technical issues related to reducing linkage error, and statistical processes to reduce the potential effects of linkage error on risk factor-outcome associations.

Chapter 8

Using all the linked resources described in chapter 6, I illustrate how these data can be used to identify population risk factors for psycho-social outcomes. I demonstrate how multiple sources of health data can be used to better characterise variables that are captured measured in several routinely collected datasets (in this case self-harm), and how they can address some of the limitations regarding identification of self-injurious behaviours, revealed during the study conducted in chapter 2. I use linked HES-CRIS and linked NPD-CRIS to build a cohort of all secondary school age individuals resident in the four south London boroughs, I use longitudinal analyses to examine the incidence of adolescent self-harm presentations to accident and emergency, and examine whether ASD is a potential risk factor for self-harm.

CHAPTER 2. CLINICAL PREDICTORS OF ANTIPSYCHOTIC USE IN CHILDREN AND ADOLESCENTS WITH AUTISM SPECTRUM DISORDERS: A HISTORICAL OPEN COHORT STUDY USING ELECTRONIC HEALTH RECORDS

The contents of this chapter have contributed to the following:

Publication in a peer-reviewed journal

Downs J, Hotopf M, Ford T, Simonoff E, Stewart R, Shetty H, Jackson R, Hayes RD. Clinical predictors of antipsychotic use in children and adolescents with autism spectrum disorders: a historical open cohort study using electronic health records. *European Child and Adolescent Psychiatry* 2016; 25: 649-658

2.1 SUMMARY

Background: Children with autism spectrum disorders (ASD) are more likely to receive antipsychotics than any other psychopharmacological medication, yet the psychiatric disorders and symptoms associated with treatment are unclear. I aimed to determine the predictors of antipsychotic use in children with ASD receiving psychiatric care.

Methods: The sample consisted of 3482 children aged 3 to 17 with an ICD-10 diagnosis of ASD referred to mental health services between 2008 and 2013. Antipsychotic use outcome, comorbid diagnoses, and other clinical covariates, including challenging behaviours were extracted from anonymised patient records.

Results: Of the 3482 children (79% male) with ASD, 348 (10%) received antipsychotic medication. The fully adjusted model indicated that comorbid diagnoses including hyperkinetic (OR 1.44, 95%CI 1.01-2.06), psychotic (5.71, 3.3-10.6), depressive (2.36, 1.37-4.09), obsessive compulsive (2.31, 1.16-4.61) and tic disorders (2.76, 1.09-6.95) were associated with antipsychotic use. In addition, clinician-rated levels of aggression, self-injurious behaviours, reduced adaptive function, and overall parental concern for their child's presenting symptoms were significant risk factors for later antipsychotic use.

Conclusions: In ASD**,** a number of comorbid psychiatric disorders are independent predictors for antipsychotic treatment, even after adjustment for familial, socio-demographic and individual factors. As current trial evidence excludes children with comorbidity, more pragmatic randomised controlled trials with long term drug monitoring are needed.

2.2 INTRODUCTION

Antipsychotics are the most common psychotropic medication prescribed to children with autism spectrum disorders (ASD).[105] US based studies suggest between 20%-34% of children with ASD receive antipsychotics.[106,107] Rates are lower in Europe, between 7-11%,[108,109] but appear to be increasing.[41] Two atypical antipsychotics in particular are most commonly used, risperidone and aripiprazole, which have been demonstrated to be effective in reducing "irritability" in children with ASD, but show limited impact on the core features of ASD.[110]

Clinicians and families face a difficult task when deciding whether antipsychotic treatment is indicated. Evidence from antipsychotic trials in childhood ASD are derived from samples that bear little resemblance to children typically seen in clinical practice, as they exclude children with formally diagnosed psychiatric comorbidity.[111] Another problem is the almost exclusive focus of trials on irritability as a target symptom in ASD. Irritability is a highly prevalent symptom in clinical settings, it has no standard taxonomy, and is associated with most childhood mental health problems.[112] Therefore, based on trial evidence, the type and severity of childhood ASD related irritability symptoms, which warrant antipsychotic treatment, are unclear. Furthermore, antipsychotic medication does not have UK marketing authorisation for use in childhood ASD, although risperidone is licenced for use in the short-term management of aggression in children with conduct disorder.[113] Balancing antipsychotic risk benefit profiles are further complicated by little safety evidence being available for children with ASD.[114,115] Antipsychotic use for children in general is associated with a number of adverse health outcomes, most commonly extrapyramidal side effects, obesity and hyperprolactinaemia.[116] Given the limited evidence base, NICE guidelines advocate cautious antipsychotic prescribing in children with ASD and only to treat severe challenging behaviours (also known as 'behaviours that challenge') such as aggression, self-injury and impulsive/ dangerous behaviours.[103]

It remains unclear how current evidence, licensing and guidance for antipsychotic use in children with ASD are applied clinically.[117] There are very few UK based naturalistic studies of prescribing in children with ASD, and, as yet, no examinations of the diagnostic predictors of antipsychotic use.[108] Comorbid psychiatric disorders are common (and frequently multiple) in children with autism spectrum disorders and may be targets for intervention.[118] Current

knowledge is largely based on parent reports in US surveys which indicate that antipsychotics are used predominantly to treat comorbid diagnoses (e.g. depression, bipolar, anxiety, conduct disorder and attention deficit hyperactivity disorders) in children with ASD.[106,107] However, these findings may not generalize to non-US clinical populations as US antipsychotic marketing,[119] prescribing policy [120,121] and practice differ markedly to the other Western Countries.[105,122] Given that the majority of the aforementioned studies report cross-sectional findings from retrospective parental accounts of both comorbidity and past medication use, the direction of effect is unclear, and recall bias may obscure true prescribing patterns. Furthermore, these studies do not account for important confounding factors, such as psychosis, adaptive function, and intellectual disability which may lead to an overestimate of the association between certain comorbidities and antipsychotic use.

To clarify how antipsychotics are used use in childhood ASD, I explored the clinical factors that predicted antipsychotic prescribing. I conducted a historical cohort study using the anonymised electronic health records of children with ASD treated by UK child and adolescent mental health services (CAMHS). As challenging behaviours (or 'behaviours that challenge') are symptoms that cut across most childhood psychopathology, I hypothesized that the common psychopathologies comorbid with ASD including hyperkinetic, oppositional and conduct, depression and anxiety disorders, would all show longitudinal associations with antipsychotic use. I also examined whether associations between comorbidity and antipsychotics were attenuated after I controlled for challenging behaviours, given that these are the most common non-psychotic symptoms formally recognized as targets for antipsychotic treatment by current national ASD management guidelines.[103]

2.3 METHODS

2.3.1 Study Setting

This study used data extracted from the anonymised, electronic clinical records of children referred to South London and Maudsley NHS Foundation Trust (SLaM) between 1st January 2008 and 31st December 2013. Over this period, the SLaM provided all aspects of specialist mental healthcare to a catchment population of approximately 280,000 children resident within four London boroughs (Lambeth, Southwark, Lewisham, Croydon). In addition to the district services, SLaM provided specialist inpatient and outpatient ASD assessment and treatment services for young people from across the UK. Each borough had a dedicated multidisciplinary service for children, which accepted referrals for school age children (4 to 18 years;

exceptionally cases are accepted below this age) with suspected or previously confirmed ASD, displaying emotional or behavioural difficulties. Children were referred from primary care, child health, and educational and social care services, and typically underwent a multidisciplinary assessment by CAMHS clinicians. Primary and secondary psychiatric disorders were diagnosed by CAMHS using the ICD-10 multi-axial classification system.[123] Semi-structured validated assessments, for example the Autism Diagnostic Observation Schedule (ADOS)[124] were used if an ASD diagnosis was unclear after initial assessment. Compared with expert consensus, there is a high specificity for ASD diagnoses by clinicians working at a district level.[125] Socio-demographic characteristics and clinical information were recorded using computerised assessment pro-forma, which included the Strengths and Difficulties Questionnaire (SDQ).[126]

The Clinical Record Interactive Search (CRIS) system

The CRIS system was used to provide an anonymised, electronic mental health records database to search on structured data and free text fields on over 35,000 child and adolescent cases referred to SLaM services. The CRIS system derives its data from the SLaM Patient Journey System (ePJS), a locally developed electronic health record (EHR) system designed to capture all clinical activity conducted by SLaM staff. Since 2007, all clinical information relating to CAMHS services have been held within ePJS. This has included risk and clinical assessment proforma, medication, clinical correspondence, progress notes, admission, discharge and outpatient appointment dates – both in structured fields , where data entry options are limited to a fixed selection of categories, and unstructured fields where data are entered as freely written text, such as a clinic letter.[127]

In 2008, the CRIS system was developed, which de-identified ePJS records by masking patient identifiers with a string of text, ZZZZZ for patients, and QQQQQ for carer identifiers. Although, this method did not completely anonymise the records – an evaluation found around 1 in 500 documents did reveal some personal identifiers [128] – it was granted NHS research ethics committee approval to conduct analyses on the de-identified data for the purposes of mental health research and audit.[129] The CRIS system was designed with robust governance structures, including a patient led oversight committee to review and minimise the risk of statistical disclosure of proposed research projects, regular patient and public facing engagement events, robust auditing, and requiring NHS research passports for researchers wishing to use the de-identified data.

The data in CRIS mirrors the information recorded in ePJS. CRIS stores these data in a relational database in different data tables. For example, there are dedicated tables for medication data, patient referral data, and progress note data. CRIS enables, patient attributes of interest (a risk or outcome variable) to be provided as a file for retrospective cohort or cross-sectional analyses. For researchers wanting to access the data, two interfaces are made available. One a web-based search engine powered by the Microsoft FAST system which allows researchers to perform *google-like* key word searches over the entire patient record. For example, a key word search of "Autism" would provide every document, nested at a patient level, that contained the term "Autism". This interface is built to enable researchers to validate data extracts, for example by reading the document from which a patient variable/ attribute has been derived. The second interface is a SQL Server Management Studio interface, which allowed users, commonly supported by CRIS informaticians, to write more complex data queries using SQL code, and extract relevant data.

In this study, Generalised Architecture for Text Engineering (GATE), a natural language processing architecture was also applied to extract data from the free text tables containing progress notes and correspondence, and combine it with structured data held within the SQL database. GATE has been comprehensively described elsewhere.[130,131] Briefly, it is a technically complex and powerful NLP tool, which allows for a variety of NLP approaches to be combined to enable key words within a document to be identified and categorised. The GATE application used in this study was designed to extract antipsychotic prescription information data from free-text, such as drug name, the status of the prescription (start date) and the date and the relative nature of the prescription (current, in the past, or planned for the future).[129]

2.3.2 Study sample

Cases were part of an open clinical cohort (entering and leaving the study at different time-points) and included children aged 3-17 years with a diagnosis of ASD (International Classification of Diseases, 10th Revision (ICD-10) F84.0, F84.1, F84.5, F84.9)[123] recorded between 1st January 2008 and 31st December 2013. Children were excluded if any past course of antipsychotic treatment was noted in the clinical record in the year prior to the observation period.

2.3.3 Measurements

Outcome: antipsychotic use

Antipsychotic use outcome data were extracted from free text fields held in the SLaM Case Register, the SLaM pharmacy dispensing database and structured medications fields in the electronic health record. Drug names listed in table 2.1, including common misspellings and abbreviations were identified using validated rules based Natural Language Processing (NLP) approaches via GATE software extraction.[132] Antipsychotic use was measured during the observation period (01/01/2008-31/12/2013).

Table 2.1 British National Formulary names used to categorise antipsychotic medication with the electronic record

Generic drug name	Trade name
Amisulpride	Solian
Aripiprazole	Abilify
Asenapine	Sycrest
Benperidol	Anquil/Benquil
Chlorpromazine	Largactil
Clozapine	Clozaril/Denzapine/Zaponex
Flupentixol/Flupenthixol	Depixol / Fluanxol
Fluphenazine	Modecate
Haloperidol	Dozic/Haldol/Serenace/Haldol
Levomepromazine	Nozinan
Olanzapine	Zyprexa / ZymAdhera
Paliperidone	Invega / Xeplion
Pericyazine/Periciazine	Neulactil
Perphenazine	Fentazin
Pimozide	Orap
Pipotiazine	Palmitate
Prochlorperazine	(Buccastem/Stemetil)
Promazine	Promazine
Quetiapine	Seroquel
Risperidone	Risperdal
Sulpiride	Dolmatil/Sulpor/Sulpitil/Sulparex
Trifluoperazine	Stelazine
Zuclopenthixol	Clopixol/Acuphase

Recorded antipsychotic use was categorised as a binary present/absent outcome. The NLP application used to extract data on antipsychotic use were validated against manual review of

300 randomly selected records resulting in precision (positive predictive value) and recall (sensitivity) statistics of 0.98 and 0.97 respectively for current use. The SLAM CAHMS decision date to prescribe antipsychotics was accurate in 89.3% to a month error margin, 95.4% within three months.

Exposure: psychiatric comorbidity & intellectual disability

The main exposure was ICD -10 recorded comorbid psychiatric diagnoses, which were extracted from free text and structured fields. ICD-10 Axis one comorbid diagnoses were categorized into: psychotic (F1x.5, F20–F29, F31, F32.3, F33.3), depressive disorders (F32), anxiety, stress and emotional (F40-41, F43-F48, F93), obsessive-compulsive (F42), hyperkinetic (F90), oppositional defiant and conduct (F91-F92) and tic (F95). Low frequency psychiatric diagnoses were collapsed into a single category labelled "Other". In addition, children were categorised according to presence of an ICD-10 Axis three diagnosis of intellectual disability (F70-F79).

Prescribing decisions were recorded contemporaneously, but there was often a short administrative lag before diagnostic reports appeared in the electronic medical record. These reports contained detailed clinical assessments conducted during the pre-medicated period. To permit inclusion of these longitudinally collected clinical data, but also ensure comorbidity exposures occurred prior to antipsychotic use, co-morbid diagnoses were only coded as present if they were recorded before, or up to 30 days after, recording of antipsychotic medication.

Covariates:

All covariates were extracted from the medical record during the initial CAMHS assessment period, and prior to antipsychotic use. Measures of challenging behaviours were taken from the SLaM risk assessment proforma. I chose assessment items with high face validity for challenging behaviours, as described by expert consensus in national guidance,[103]including physical aggression against self (self-injury), violence and aggression to others or property (aggression), harm through loss of self-care such as not drinking or eating (self-neglect), impulsive and dangerous acts (high-risk behaviours), and habitual behaviours related to intellectual disability such as rocking or skin picking that can cause injury (ID related harm). Clinicians rated severity along a 4-point categorical scale: 'None', 'Low', 'Moderate', or 'High'. For ease of clinical interpretation and to ensure adequate numbers in each category, this scale was converted into a binary variable for each behaviour domain (Moderate or High rating categorised as High risk=1; None or Low ratings, low risk=0). Children's adaptive

functioning was rated using Children's Global Assessment Scale (CGAS) [133], except for those children with significant ID, where the Developmental Disabilities-CGAS was used in some cases.[134] Higher scores (range 0-100) are associated with better functioning.

Demographic and family covariates consisted of gender, ethnicity, history of parental mental illness, and clinician-rated levels of parental concern for their child's symptoms at their initial presentation to CAMHS, which were retrieved using CRIS from structured fields in the source dataset. Age at CAMHS assessment was calculated at the date of the first recorded diagnosis of autism spectrum disorder within the clinical record. UK Census data provided small area (average 400 households) level deprivation scores.[135]

Emotional, hyperactive and conduct problem domains were assessed via the caregiver versions of the 25-item Strengths and Difficulties Questionnaire which has sound psychometric properties in clinical samples.[136] These were available in the clinical record for a third of the sample (n=1234, 35%).

2.3.4 Analysis

To authenticate clinically-recorded common comorbid conditions (depression and anxiety, hyperkinetic and conduct disorders) I used a sub-group of the cohort with completed SDQs (n=1234). Independent sample t-tests were used to test for statistical differences in parental reported SDQ psychopathology subscale scores (emotional, hyperactivity, conduct problem subscales) for children with and without these common comorbid conditions.

Logistic regression was used to examine whether antipsychotic use was predicted by demographic characteristics, psychiatric comorbidities, intellectual disability, adaptive function, behaviours that challenge, parental characteristics and neighbourhood deprivation. Multivariable analyses were then conducted to examine the effect of each of these variables on antipsychotic use after adjusting all other individual and contextual covariates (listed in table 2.2). The following sensitivity analyses were carried out: i) using non-aggregated challenging behaviours categories (4 levels) as the binary variable may have introduced residual confounding; ii) excluding those who came from outside the local catchment (these individuals may have had substantial contact with non-SLaM services not represented in the CRIS source dataset); iii) selecting those with only one comorbid disorder and modelling the effect of a

single comorbidity on antipsychotic use (without adjusting for the full set of covariates). All analyses were conducted using Stata version 12.

2.4 RESULTS

2.4.1 Demographic and clinical characteristics of the sample

Over the six-year observation period, I identified 3482 children aged below 18 years (2686 male and 796 female) with a diagnosis of ASD. The mean (SD) exposure to child mental health services, defined as the time between the date of recorded ASD diagnosis and the end of the observation period or date of 18th Birthday (whichever sooner) was 968 (597) days. Three hundred and forty-eight children were prescribed antipsychotics, mainly risperidone (55%, n=191) and aripiprazole (32%, n=112). All were receiving adjunctive psycho-social interventions. Table 2.2 shows the characteristics of the total sample and those prescribed antipsychotics. Nearly 75% of children prescribed antipsychotics were in the adolescent age range (age: 13-18 years), representing a 6-fold risk (OR 6.29, 95% C.I 3.40-12.1) relative to early childhood (age: 3-6 years).

ICD -10 recorded comorbid psychiatric diagnoses were present in 54% of the sample, a quarter diagnosed with a hyperkinetic disorder, and 20% diagnosed with intellectual disability. Table 2.3 provides further details of comorbid psychiatric diagnoses by antipsychotic use. Antipsychotics were prescribed to approximately half of children with ASD and psychotic disorder, and over a quarter of children diagnosed with obsessive compulsive disorder, or tic disorders.

Table 2.2 Individual and contextual characteristics of 3482 children with autism spectrum disorders and antipsychotic use referred to local and specialist Child and Adolescent Mental Health Services.

	Total sample (n=3482)		Sample receiving antipsychotics (n=348)	
Male	2686 (77.1%)		249 (71.6%)	
Female	796 (22.9%)		99 (28.4%)	
Child age category	At CAMHS assessment		At antipsychotic use	
	n (%)	mean (s.d)	n (%)	mean (s.d)
Early(3-6yrs)	362 (10.4%)	4.9 (0.77)	3 (0.9%)	5.7 (0.3)
Mid (6-12 yrs)	1664 (47.8%)	9.0 (1.69)	89 (25.6%)	9.6 (1.4)
Late (13-17 yrs)	1456 (41.8%)	14.8 (1.66)	256 (73.5%)	15.1 (1.6)
Ethnicity				
White British	1683 (48.3%)		208 (59.8%)	
White Other	170 (4.9%)		11(3.2%)	
East Asian	68 (2.0%)		9 (2.6%)	
British/ Black African	651 (18.7%)		57 (16.4%)	
British/ Black Caribbean	130 (3.7%)		5 (1.4%)	
Mixed Heritage	386 (11.1%)		37 (10.6%)	
South Asian	84 (2.4%)		11 (3.2%)	
Not Stated	310 (8.9%)		10 (2.8%)	
Adaptive function: Children's Global Assessment Scale (CGAS)[a]				
0-25 (most impaired)	174 (5.7%)		50 (15.0%)	
25-50	1346 (43.8%)		219 (65.6%)	
50-75	1465 (47.7%)		62 (18.5%)	
75-100	89 (2.90%)		3 (0.9%)	
Challenging Behaviours	474 (16.7%)		129 (43.0%)	
Self injury[b]				
ID related harm[c]	1097 (40.7%)		150 (51.4%)	
Aggression[d]	1165 (40.3%)		210 (67.7%)	
Self neglect[e]	300 (10.5%)		77 (25.0%)	
High risk behaviours[f]	654 (23.1%)		140 (46.5%)	
Family Characteristics				
Caregiver mental illness[g]	685 (22.9%)		73 (22.8%)	
Caregiver Substance Misuse [g]	201 (6.7%)		16 (5.0%)	
Parental Concern[h]	2033 (69.9%)		278 (89.4%)	
Neighbourhood Characteristics[i]				
Level of Deprivation (Tertiles)				
1st (least deprived)	1064 (32.8%)		142 (44.4%)	
2nd	1093 (33.7%)		95 (29.7%)	
3rd (Most Deprived)	1089 (33.6%)		83 (25.9%)	

* Missing cases = [a]408 [b]640 [c]785 [d]593 [e]620 [f]653 [g]488 [h]573 [i]236

Table 2.3 Prevalence of comorbid psychiatric disorder and antipsychotic treatment in 3482 children with autism spectrum disorders

ICD-10 Disorder	Total sample (n=3482)	Sample receiving antipsychotics (n=348)
Any comorbid disorder	1897 (54.5%)	285 (81.9%)
Hyperkinetic	862 (24.8%)	121 (34.8%)
Oppositional and Conduct	256 (7.3%)	51 (14.7%)
Depression	154 (4.4%)	36 (10.3%)
Anxiety, Emotional and Stress	279 (8.0%)	45 (12.9%)
Obsessive Compulsive	97 (2.8%)	26 (7.5%)
Tic	51 (1.5%)	13 (3.7%)
Psychosis	116 (3.3%)	54 (15.5%)
Intellectual Disability	656 (18.8%)	114 (32.8%)
Other **	129 (3.7%)	18 (5.2%)

** remaining, rarely occurring diagnoses, were collapsed into a single category labelled Other (includes ICD-10 F50 eating disorders, F04-09 organic disorders, F1x.1-4 substance misuse, F94.1-2 attachment disorders)

2.4.2 Authentication of co-morbid diagnoses against the SDQ

In the authentication analyses, I found that ASD children diagnosed with comorbid emotional (depression and anxiety), hyperkinetic or conduct disorders had significantly higher SDQ subscales scores within their respective SDQ domains (emotional, hyperactive, conduct) compared with children without the respective comorbid diagnosis (see table 2.4). This sub-sample (n=1234) were broadly representative with the remaining cohort. Male gender (77.9% vs 76.8%) mean age at recorded ASD diagnosis (10.5 vs 11.2 years), White British (43.5% vs 46.3%) SDQ relevant clinical diagnoses: Emotional disorder (15.1% vs 12.8%) and Hyperkinetic (23.2% vs 27.2%), Conduct (6.4% vs 7.9%); and antipsychotic use (9.4% vs 10.3%).

Table 2.4 Comorbid disorders diagnosed by clinicians and validated against parental Strength and Difficulties Questionnaire subscale score in sub-sample of children with ASD (n=1234)

Clinical Diagnoses		SDQ subscale, n=1234 (mean, SD)		
		Emotional	Conduct	Hyperactivity
Depressive and Emotional Disorders				
Depressive disorders (F32), Anxiety, stress and emotional (F40-41, F43-F48, F93), Obsessive-compulsive (F42)	Present (n=171)	6.5(2.8)[a]	3.72(2.5)	6.35(2.8)
	Absent(n=1063)	4.6(2.5)[a]	4.25(2.5)	7.36(2.5)
Externalizing Disorders				
Oppositional / Conduct Disorders (F91-F92)	Present (n=81)	5.27(2.7)	6.0(2.5)[b]	7.61(2.3)
	Absent(n=1153)	4.85(2.8)	4.14(2.5)[b]	7.19(2.6)
ADHD (Hyperkinetic, F90)	Present (n=345)	4.52(2.8)	5.24(2.5)	8.44(1.9)[c]
	Absent(n=889)	5.02(2.8)	3.88(2.4)	6.74(2.7)[c]

[a] t =8.57, p<0.001 [b] t = 6.5, p<0.001 [c] t =10.8, P<0.001

2.4.3 Socio-demographic and clinical factors and their associations with antipsychotic treatment

In the adjusted model, positive associations with antipsychotic use remained significant for age at the time of assessment, clinician-rated aggression, self-injurious behaviour, and high parental concern for their child's symptoms at initial presentation (see table 2.5). In addition, adaptive function and the presence of caregiver substance misuse showed strong inverse associations with antipsychotic use. Associations with ethnicity, caregiver mental illness and neighbourhood deprivation were non-significant.

Table 2.5 Multivariable model of antipsychotic use in children with ASD by socio-demographic characteristics and other covariates

Patient characteristics	O.R (95% CI) (n=3482)	*P*	aO.R (95% CI)*	*P*
Female sex (vs male)	1.39 (1.09-1.79)	0.009	1.02(0.71-1.46)	0.89
Age at CAMHS assessment	1.18 (1.15-1.23)	<0.0001	1.11(1.05-1.16)	<0.001
Ethnicity				
White British	reference			
White Other	0.49(0.26-0.92)	0.026	0.62(0.24-1.55)	0.31
East Asian	1.08(0.52-2.21)	0.83	0.91(0.35-2.31)	0.84
British/ Black African	0.68(0.50-0.92)	0.014	1.14(0.73-1.78)	0.55
British/ Black Caribbean	0.28(0.11-0.70)	0.006	0.56(0.21-1.54)	0.26
Mixed Heritage	0.75(0.52-1.09)	0.13	0.78(0.46-1.32)	0.36
South Asian	1.06(0.56-2.05)	0.84	1.26(0.47-3.32)	0.70
Not stated	0.24(0.12-0.45)	<0.001	0.27(0.09-0.80)	0.02
Adaptive function: Children's Global Assessment Score (CGAS)	0.95(0.94-0.95)	<0.0001	0.96(0.95-0.97)	<0.0001
Challenging Behaviours				
Self-injury	4.80(3.72-6.20)	<0.0001	1.85(1.30-2.63)	<0.0001
ID related harm	1.63(1.27-2.07)	<0.0001	0.72(0.49-1.06)	0.10
Aggression	3.57(2.77-4.59)	<0.0001	2.14(1.50-2.06)	<0.0001
Self-neglect	3.48(2.61-4.67)	<0.0001	1.20(0.78-1.80)	0.35
High risk behaviours	3.40(2.66-4.35)	<0.0001	1.22(0.86-1.73)	0.27
Family Characteristics				
Caregiver mental illness	1.0(0.75-1.31)	0.98	0.87(0.60-1.26)	0.47
Caregiver Substance Misuse	0.71(0.42-1.20)	0.20	0.57(0.30-1.08)	0.09
High Parental Concern	4.05(2.79-5.85)	<0.0001	2.02 (1.27-3.22)	0.003
Neighbourhood Characteristics[i]				
1st (least deprived)	reference			
2nd	0.62(0.47-0.81)	0.001	0.82(0.56-1.19)	0.31
3rd (Most Deprived)	0.54(0.40-0.71)	<0.0001	0.91(0.62-1.35)	0.65

*aO.R, adjusted Odds Ratio, adjusting for socio-demographic and parental and neighbourhood characteristics, challenging behaviours, adaptive function and co-existing ICD-10 Mental and behavioural disorder groupings : hyperkinetic (F90), depressive disorders (F32), psychosis (F1x.5, F20–F29, F31, F32.3, F33.3), Oppositional and Conduct (F91-F92), anxiety, stress and emotional (F40-41, F43-F48, F93), Obsessive-compulsive (OCD, F42), Tic (F95), Intellectual Disability (ID, F70-F79) and Other psychiatric diagnosis.

Table 2.6 shows that a number of comorbid ICD-10 mental disorders, even after adjustment for all other covariates and comorbidities, remained significantly associated with antipsychotic use including hyperkinetic (OR 1.44, 95%CI 1.01-2·06), psychotic (OR 5.71, 3.3-10.6), depressive (2.36, 1.37-4.09), obsessive compulsive (2.31, 1.16-4.61) and tic disorders (2.76, 1.09-6.95). These associations remained when antipsychotic use was compared between ASD children with no-comorbidity with those who only had the specific comorbidity alone, rather than multiple comorbidities (for example, only comorbid hyperkinetic disorder, see table 2.7).

Table 2.6 Multivariable model of antipsychotic use in a cohort of children with ASD by psychiatric comorbidity (n=3482)

ICD-10 Disorder	O.R (95% CI)	*P*	aO.R* (95% CI)	*P*
Any comorbid disorder	4.27(3.22-5.66)	<0.0001	-	
Hyperkinetic	1.73(1.36-2.18)	<0.0001	1.44(1.01-2.06)	0.042
Oppositional and Conduct	2.47(1.77-3.43)	<0.0001	1.55(0.96-2.51)	0.073
Depression	2.95(1.99-4.36)	<0.0001	2.36(1.37-4.09)	0.002
Anxiety, Emotional and Stress	1.84 (1.31-2.59)	<0.0001	1.20(0.72-1.98)	0.484
Obsessive Compulsive	3.48(2.19-5.53)	<0.0001	2.31(1.16-4.61)	0.017
Tic	3.16(1.67-5.99)	<0.0001	2.76(1.09-6.95)	0.032
Psychosis	9.1 (6.19-13.4)	<0.0001	5.71(3.28-10.6)	<0.0001
Intellectual Disability	2.33(1.82-2.97)	<0.0001	1.68(1.11-2.53)	0.015
Other **	1.49 (0.89-2.48)	0.13	1.62(0.83-3.16)	0.157

*aO.R, adjusted Odds Ratio, adjusting for socio-demographic and parental and neighbourhood characteristics, challenging behaviours, adaptive function and co-existing ICD-10 Mental and behavioural disorder groupings : hyperkinetic (F90), depressive disorders (F32), psychosis (F1x.5, F20–F29, F31, F32.3, F33.3), Oppositional and Conduct (F91-F92), anxiety, stress and emotional (F40-41, F43-F48, F93), Obsessive-compulsive (OCD, F42), Tic (F95), Intellectual Disability (ID, F70-F79) and Other psychiatric diagnosis.

** remaining, rarely occurring diagnoses, were collapsed into a single category labelled Other (includes ICD-10 F50 eating disorders, F04-09 organic disorders, F1x.1-4 substance misuse, F94.1-2 attachment disorders)

2.4.4 Sensitivity analysis

Specified sensitivity analyses that used non-aggregated behaviour categories produced little change to the overall pattern of results in the fully adjusted models, with the direction and magnitude of effect being consistent across the comorbidities. Similarly, removing the children resident outside the local catchment area from the sample (n=1170, 33%) produced little change, with the exception that oppositional defiant and conduct disorder, produced imprecise estimates related to the very low number of children prescribed antipsychotics.

Table 2.7 A comparison of antipsychotic treatment between children with no comorbidity and singleton comorbid disorder only in Autism Spectrum Disorders.

ICD-10 Disorder	No antipsychotics n (%)	Received antipsychotics n (%)	O.R (95% C.I.)	*P*
ASD (no comorbid disorder)	1522 (96.0)	63(4.0)	reference	
Singleton comorbid disorder				
Hyperkinetic	454 (91.4)	43 (8.7)	2.29 (1.53-3.41)	<0.0001
Oppositional and Conduct	73 (93.6)	5 (6.4)	1.65 (0.64-4.23)	0.29
Depression	59 (81.9)	13 (18.1)	5.32 (2.77-10.2)	<0.0001
Anxiety, Emotional and Stress	136 (95.1)	7 (4.9)	1.23 (0.55-2.77)	0.59
Obsessive Compulsive	51 (82.3)	11 (17.7)	5.21 (2.59-10.5)	<0.0001
Tic	20 (86.9)	3 (13.0)	3.61 (1.04-12.5)	0.03
Psychosis	17 (42.5)	23 (57.5)	32.7 (16.6-64.2)	<0.0001
Intellectual Disability	329 (85.7)	55 (14.3)	4.04 (2.75-5.91)	<0.0001
Other **	43 (95.6)	2 (4.4)	1.12 (0.27-4.74)	0.87

*aOR, adjusted Odds Ratio, adjusting for socio-demographic and parental and neighbourhood characteristics, challenging behaviours, adaptive function and co-existing ICD-10 Mental and behavioural disorder groupings : hyperkinetic (F90), depressive disorders (F32), psychosis (F1x.5, F20–F29, F31, F32.3, F33.3), Oppositional and Conduct (F91-F92), anxiety, stress and emotional (F40-41, F43-F48, F93), Obsessive-compulsive (OCD, F42), Tic (F95), Intellectual Disability (ID, F70-F79) and Other psychiatric diagnosis.

** remaining, rarely occurring diagnoses, were collapsed into a single category labelled Other (includes ICD-10 F50 eating disorders, F04-09 organic disorders, F1x.1-4 substance misuse, F94.1-2 attachment disorders)

2.5 DISCUSSION

In the largest study to date using non-administrative, clinical mental health records in ASD, I found antipsychotic prescribing for children with ASD was strongly associated with comorbidity. Intellectual disability and psychiatric comorbidities, including hyperkinetic, depression, psychotic, obsessive compulsive and tic disorders, were all associated with antipsychotic treatment, even after controlling for clinician-rated challenging behaviour symptoms at initial assessment. I also found increasing age, aggression, self-injurious behaviour, level of adaptive function, and parental concern were all significant predictors of antipsychotic use.

The observed association between antipsychotic use and age is consistent with previous ASD studies.[106,137] Over two-thirds of children treated with antipsychotics were adolescents. This highlights the need for more trials that include this age group but also suggests that treatment acceptability, and hence trial recruitment, will be more feasible than in younger children. Social factors also appeared to play a role; clinicians who perceived greater parental concern for children's presenting symptoms were more likely to prescribe antipsychotic treatment. I was not aware of any prior studies that measure parental influences on antipsychotic use in ASD, however the study finding is consistent with previous work that show a positive association between parental strain and medication treatment for childhood disruptive disorders.[138] Consistent with several other investigations in clinical samples,[139,140] the unadjusted analyses suggests that there may be discrepancies between ethnic groups regarding prescribing antipsychotic medication to children. However, in keeping with a more recent study on psychotropic use in children, I found that after adjustment for markers of clinical severity, ethnicity was no longer significantly associated with antipsychotic use.[141]

Using a historical cohort design in a clinical sample of children with ASD, this is the first longitudinal study of challenging behaviours and psychiatric comorbidity profiles as predictors of antipsychotic use. The results suggest that clinicians are using antipsychotics where they are known to be efficacious;[110] to target aggression and self-injurious behaviours. Many studies so far have been hindered by parental report of comorbidities and medication use, retrospective or cross sectional design, or the confounding effect of unmeasured psychiatric symptoms and disorder severity not being accounted for.[106,107,142] In addressing these limitations I found that, unlike a number of US studies, antipsychotics were not significantly associated with comorbid emotional disorders.[106,107,142]

2.5.1 Strengths

This study has a number of strengths: I used longitudinally collected clinician recorded data in an unselected population of children and adolescents with ASD referred to CAMHS to study off-label antipsychotic use. This avoids the non-response or recall bias issues that may arise in surveys of parents. The sample included the entire psychiatric population of four south London boroughs for school age children (4 to 18 years) with suspected or previously confirmed ASD and displaying emotional or behavioural difficulties, in addition to children from other areas of the UK referred to National & Specialist services. However, because I studied a cohort enriched by national referrals, the prevalence of psychiatric comorbidity and antipsychotic treatment should not be taken as representative of the children with ASD in the general population.

2.5.2 Limitations

This study has limitations. First, an ASD diagnosis may 'overshadow' other psychiatric diagnoses and reduce the likelihood of clinicians recording additional psychiatric diagnoses. For example, ICD-10 criteria preclude the diagnosis of hyperkinetic disorder being given once ASD is established, which may lead to an underestimate of the association between hyperkinetic comorbidity and antipsychotic use. That said, many clinicians override this instruction based on recent evidence from clinical and treatment studies. Second, the type of assessment and treatments offered to families may vary by clinician. In the analysis, I lacked detailed information about the assessing and prescribing clinician and could not account for variation in practice. Third, I did not include physical comorbidities (e.g. epilepsy, obesity), other pharmacological treatments or duration of psycho-social interventions which may act as potential confounders to antipsychotic use. Fourth, I did not apply a research scale to measure challenging behaviours,[143] and the items I used lacked key contextual information, for example the timing, frequency or intentionality of the challenging behaviour – when was the self-injury conducted, was there wilful intent to self-harm, was it conducted with suicidal intention? Instead, I used risk assessment items commonly mandated for use in clinical mental health services,[144] which could likely aid study replication in other UK settings. Fifth, I coded comorbid disorders preceding and up to 30 days post antipsychotic use, which prevented the exclusion of pre-medication diagnostic reports. Theoretically, this could introduce an observer bias, as the intensity of observation by the CAMHS service post antipsychotic treatment may increase a child's risk of having a clinically recorded comorbid disorder. However, iatrogenic comorbid psychiatric conditions are very unlikely to develop or be recorded within this short

timeframe. Last, due to limitations in the free text coding and extraction process, I cannot exclude residual confounding as an influence on the findings, especially within the broad diagnostic categories of psychotic disorder or intellectual disability. I was unable to accurately categorise the degree of intellectual disability, nor characterise the severity or duration of psychotic disorder from the electronic health records. However, I did address potential confounding due to severity of psychotic disorders and intellectual disability to some extent by the inclusion of Children's Global Assessment Score as a covariate in the final multivariable models. Residual confounding may remain nonetheless.

The findings reflect the complexity of assessing and treating comorbid psychiatric disorders in ASD. For example, ASD and psychotic disorders pose a common diagnostic challenge to clinicians given their overlapping characteristics and high potential for co-occurrence.[145,146] I found only 47% of children with ASD and psychosis received antipsychotics. This low treatment rate may be due diagnostic uncertainty. Children with ASD may be more likely to have their diagnosis of psychosis withdrawn after further clinical assessment, and before the initiation of antipsychotic treatment. A second reason may relate to clinicians, children and their families deciding that some psychotic symptoms in ASD do not warrant antipsychotic treatment. Evidence that may dissuade those from starting antipsychotic treatment include findings from longitudinal studies, which show fluctuating psychotic symptoms in children with features of autism can have a relatively benign course.[147,148]

2.5.3 Conclusions

The study findings provide a detailed account of current antipsychotic prescribing practices in a clinical population of children with ASD, which show that aggression and self-injurious behaviours are significantly associated with antipsychotic use. Irritability may be an underlying treatment target driving the association between these behaviours and antipsychotic treatment. It may also underlie the associations I found between antipsychotic use and hyperkinetic, depressive and obsessive-compulsive disorders. Alternatively, disorder specific symptoms may be targeted. For example trial data has shown risperidone and aripiprazole both significantly reduce hyperactivity and obsessional compulsive symptoms in ASD.[143,149] The study findings highlight the need for further research in childhood ASD to determine which psychotic phenomena warrant antipsychotic medication. This would help clinicians reduce the harms associated with both antipsychotic underuse (i.e. prolonging the duration of untreated psychosis) and overuse. Future research might valuably include children without ASD as comparison groups and employ more intricate text extraction methodologies to assess symptom

specific severity and impairment. This would reduce the residual confounding effects that may occur when using broad diagnostic categories, and determine whether comorbid psychiatric diagnoses in clinical practice are approached differently in children with ASD.

The findings highlight a mismatch between current clinical trials and the evidence needed to support clinical practice in ASD. Antipsychotic use was much greater in adolescents and for those with comorbid diagnoses. However, most published trials exclude children with comorbidity and rarely recruit adolescents.[110,111,114] Importantly, I show social factors play a significant part in antipsychotic use. This provides an impetus to examine the association of antipsychotic treatment against contextual, as well clinical factors. Controversy between the potential harm of both over- and under-use of antipsychotics in children with ASD continues, and underlies considerable public concern.[117] Large scale cohort studies in real world settings, such as ours, eventually leading to pragmatic trials using electronic patient records, will help this debate become better informed.

CHAPTER 3. DETECTION OF SUICIDALITY IN ADOLESCENTS WITH AUTISM SPECTRUM DISORDERS: DEVELOPING A NATURAL LANGUAGE PROCESSING APPROACH FOR USE IN ELECTRONIC HEALTH RECORDS

Publication in a peer-reviewed journal

Downs J, Velupillai S, Gkotsis G, Holden R, Kikoler M, Dean H, Fernandes A, Dutta R. Detection of Suicidality in Adolescents with ASD: Developing a Natural Language Processing Approach for Use in Electronic Health Records. *Proceedings of the American Medical Informatics Association*. (in press)

3.1 SUMMARY

Background: It is estimated that 15% of young people with ASD will contemplate or attempt suicide during adolescence, making them 30 times more at risk than typically developing children. However, there are very few epidemiological investigations of suicidality in young people with ASD, and current studies are based on small samples and subject to a number of methodological weaknesses. Electronic health records (EHRs) can be used to create retrospective clinical cohort data for large samples of children with ASD. However, systems to accurately extract suicidality-related concepts need to be developed so that putative models of suicide risk in ASD can be explored.

Methods: I present a systematic approach to 1) adapt Natural Language Processing (NLP) solutions to screen with high sensitivity for reference to suicidal constructs in a large clinical ASD EHR corpus (230,465 documents), and 2) evaluate within a screened subset of 500 patients, the performance of an NLP classification tool for positive and negated suicidal mentions within clinical text.

Results: When evaluated, the NLP classification tool showed high system performance for positive suicidality with precision, recall, and F1 scores all > 0.85 at a document and patient level.

Conclusions: The application provides accurate output for epidemiological research into the factors contributing to the onset and recurrence of suicidality, and potential utility within clinical settings as an automated surveillance tool.

3.2 INTRODUCTION

Recent studies report that over 1 in 6 young people with ASD will contemplate or attempt suicide during childhood, making them 30 times more at risk than typically developing children.[150,151] Why children with ASD have higher rates of suicidal behaviours is unclear. It is possible that risk factors for childhood suicidal behaviour found in typically developing children, such as depression or being bullied, are more prevalent or potentially have a greater negative impact in children with ASD.[150] However, very little work has been conducted in ASD cohorts, and findings derived from non-ASD samples cannot be assumed to generalise to children with ASD.[152] A growing number of studies have shown that putative risk factors (both environmental and genetic) for psychiatric outcomes can have different effects in children with neurodevelopmental disorders.[118,153] Therefore individuals with ASD may express and manifest suicidal tendencies and behaviours in ways that differ from those observed in typical development.[154]

Given the widespread adoption of Electronic Health Records (EHRs) in primary and hospital care systems and the rapid growth of health informatics capabilities, longitudinal data from large samples of children with ASD can be used to develop and test new models of suicide risk behaviour. There is considerable potential to adapt EHR research methodologies used in recent epidemiological and risk factor studies[155] and apply these approaches to address the evidence gap in ASD and other vulnerable adolescent groups.[156] Although to capitalise on these developments for suicide research, accurate EHR data extraction systems need to be developed to capture data on those young people with ASD who present to public health services with suicidal thoughts or behaviours.

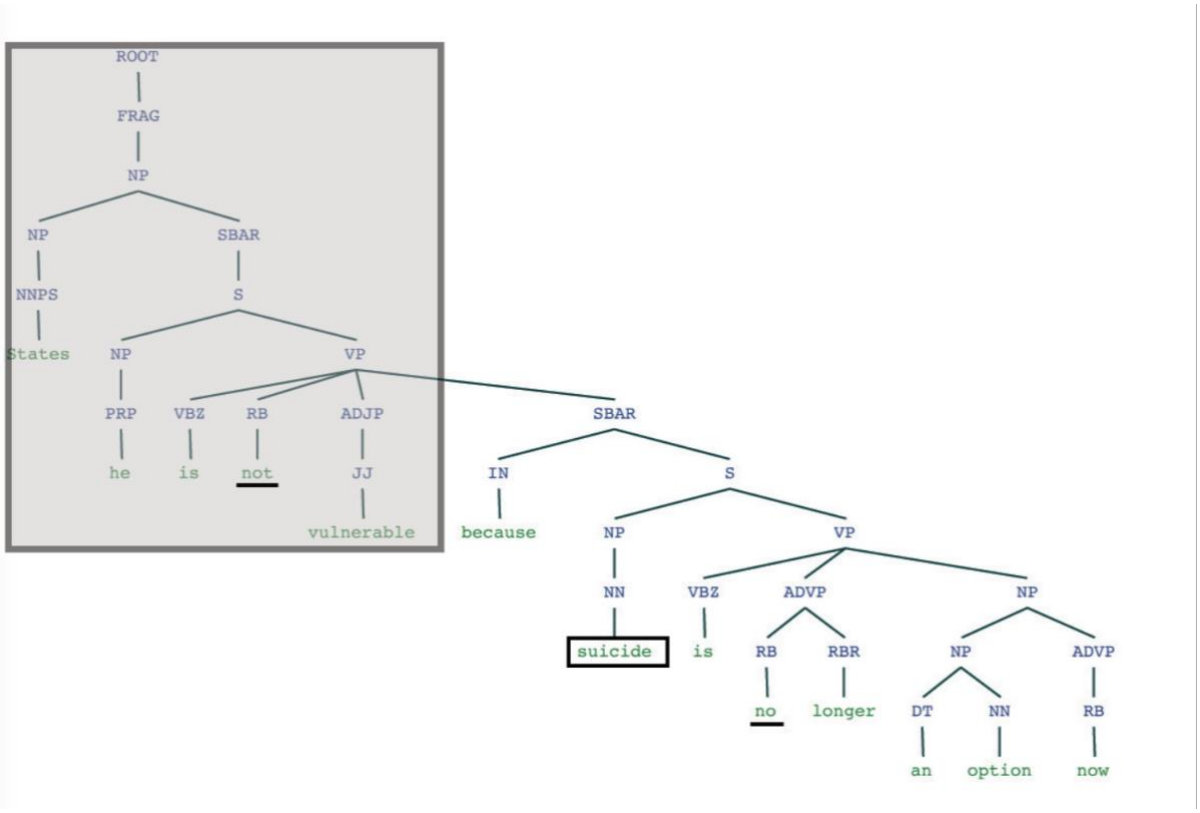
Information about suicidality in clinical documents is predominantly written in free-text. Haerian et al. showed that using only ICD-9 E-codes to detect patient-level suicide and suicide ideation from clinical text had the lowest positive predictive value (PPV): 0.55, while a combination of codes and Natural Language Processing (NLP) had the highest: 0.97, when applied on EHRs from the New York Presbyterian Hospital/Columbia University Medical Center.[157] They used MedLEE (Medical Language Extraction and Encoding System)[158] to generate Concept Unique Identifiers (CUIs) related to suicidality, and to filter out negated

mentions as well as mentions not related to the patient. Anderson et al. applied a rule-based NLP approach to identify positive or negated mentions related to suicidality in the History of Present Illness (HPI) section of EHRs from a distributed health network of primary care organizations in the US, and found that suicidality information was predominantly recorded in free-text.[94]

Because suicidality is routinely assessed in mental health care, the absence or negation of suicidal behaviour is also documented in EHRs. An NLP tool developed specifically for detecting negated mentions of suicide in mental health records using syntactic tree information was developed for use in mental health records with high accuracy (91.9%) when evaluated on 6,000 sentences from mental health EHRs. This tool has been extensively described elsewhere.[159] In brief, the tool worked by first organising the terms within each free sentence, which contained the *suicid** target word (the * indicates a wildcard permitting different suffixes on *suicid*), into a data structure, called a constituency-based parse tree structure. The tool labelled the components of the sentence to fit a root-branch-leaf organisation. This method of organising language is derived from linguistic theory of Latin and Greek grammars, so that every sentence (the root) is branched into subject (noun phrase, NP) and predicate (verb phrase, VP), and then branched further into other syntactic categories. After parsing, the NLP tool classified each target mention in the text (e.g. *suicid**) as negated or positive using a set of 15 negation terms and *pruning* rules applied to the structured sentence. Figure 3.1, taken from Gkotsis et al.,[159] gives an example of how a sentence containing the word *suicid** may be parsed into a tree structure, with the appropriate fragments selected for affirmation or negation of patient suicidality.

The aim of this study was to extend, further develop and robustly evaluate a NLP approach which could accurately identify suicidality in ASD-patients' clinical records, with the future goal that it may provide data to enable improved risk prediction for related major adverse events, such as suicide attempts.

Figure 3.1 The data structure of a sentence with a target suicide term. The constituency-based parse tree and negation rules prune fragments of the sentenced to permit accurate classification (taken from Gkotsis et al.[159])



Note: Parts of speech are tagged as Noun Phrase (NP), Proper Noun Plural Form (NNPS), subordinate clause (SBAR), Sentence (S) Verb Phrase (VP), Verb (VBZ), Adverb (RB), Adjective (ADJP) Determiner (DT), NN (noun), Preposition (IN), Fragment (FRAG)

Using EHR documents, such as progress notes, risk assessments and medical correspondence, I examined whether negation detection methods could be used to accurately identify references to suicidality in the EHRs of adolescents with ASD presenting to clinical mental health services. I defined suicidality as either the reporting of the intention to engage in a potentially lethal act towards oneself, or undertaking such acts themselves. To achieve the study aim, I developed coding rules using expert consensus, to define explicit suicidality-related mentions for adolescents with ASD seen in specialist mental health clinics (inpatient and ambulatory). Based on these rules I extended the NLP tool to 1) identify documents containing suicide-related (SR) information (i.e. NLP tool to screen documents) and 2) identify positive and negated references of suicidality on a document and patient level [i.e. NLP to classify SR documents and patients as positive, (SR-Pos), or negative, (SR-Neg)] across a large number of

EHRs. I then compared the performance of the NLP tool against expert human-rater case note reviews.

3.3 MATERIALS AND METHODS

3.3.1 Data resources

This study used data extracted from the anonymised, electronic clinical records of a sample of adolescents with ASD referred to SLaM. This sample and clinical setting has been described previously in chapter 2,[160] but in brief SLaM provides specialist inpatient and outpatient ASD assessment and treatment services for young people from across the UK. Children and adolescents in this study were referred from primary care, child health, and educational and social care services, and typically underwent a multidisciplinary assessment by Child and Adolescent Mental Health Service (CAMHS) clinicians. Primary and secondary psychiatric disorders were diagnosed by CAMHS using the International Classification of Diseases, 10th Revision (ICD-10) multi-axial classification system.

The CRIS system[129] (see chapter 2 for more details) was used to produce an anonymised EHR dataset to search on structured data and free text fields for all ASD patients. The patients were part of an open clinical cohort (entering and leaving the study at different time points) and included children aged 3–17 years with a diagnosis of ASD (ICD-10 F84.0, F84.1, F84.5, F84.9) recorded between 1 January 2008 and 31 December 2013. Free text entries, correspondence and reports were available for this sample from their initial assessment until June 2016. The resulting cohort contained 3,642 unique patients (complete age range). For the purposes of this study, I selected from an adolescent sub-sample who had at least one contact with CAMHS (i.e. one free text document in CRIS) between the ages 14 and 18 years, totalling 1,906 patients.

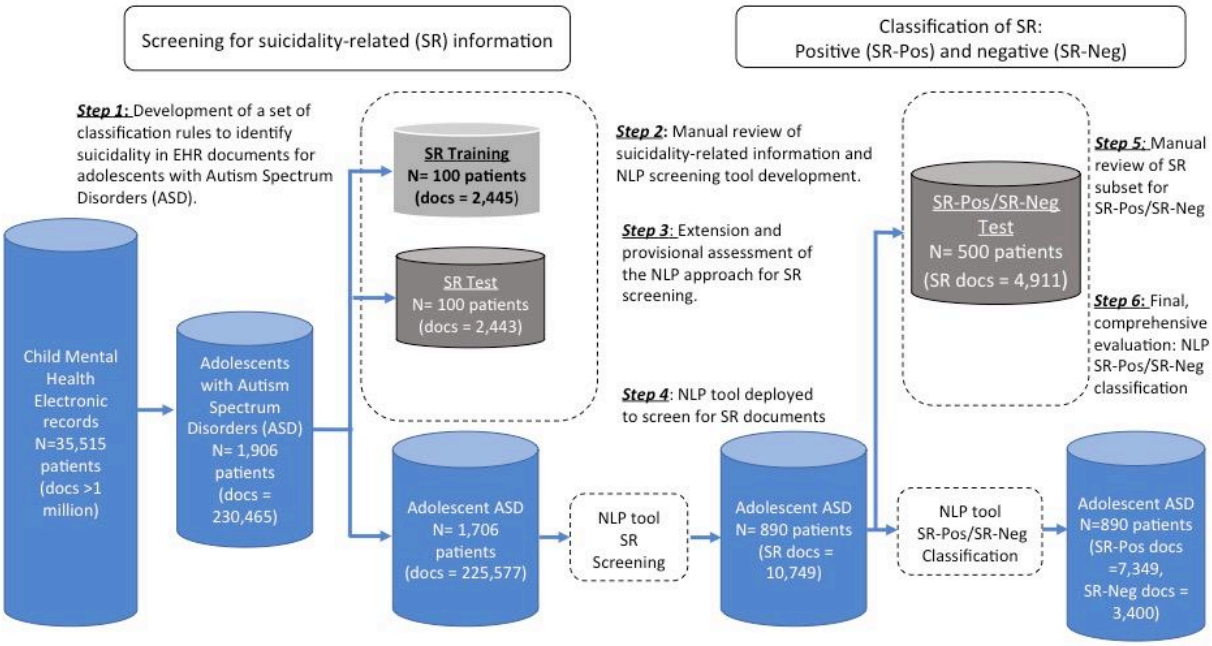
3.3.2 Overall workflow

Figure 3.2 outlines the overall workflow of the study. There were three main phases. The first phase related to the definition of classification rules to identify suicidality-related information in EHR documents for adolescents with ASD (step 1 below). These rules were then applied in the second phase where a manual review of documents (step 2) was used to inform the development and evaluation of the NLP approach to screen for SR mentions in documents and filtering out documents with no mentions related to suicidality (NSR) – step 3 below. The NLP approach was then used to extract SR documents for the third phase (step 4). In the third phase, a manual review of documents was performed to annotate mentions of suicidality in SR documents as positive (SR-Pos), negative (SR-Neg) or uncertain (SR-U), step 5. Finally, the NLP approach was evaluated for its ability to correctly classify SR-Pos or SR-Neg in these documents and patients, step 6.

Step 1: Development of a set of classification rules to identify suicidality in adolescents with ASD.

As part of a group of senior clinicians with expertise in the clinical management of neurodevelopmental disorders and suicidality assessment, I developed a set of rules to classify explicit mentions of suicidality in every document as either positive, negated or unknown. Positive mentions included text that referred to previous attempts, the presence of current or past plans of suicidal acts, command hallucinations related to carrying out a suicide attempt, a desire to be dead, researching suicide methods, having ideas or describing plans of how to end their life or, a clinical opinion of the young person being at an elevated risk of attempting suicide. Negated terms included clinical opinions of the young person **not** being at elevated risk of suicide, and recorded denial of suicidality by the young person (either directly or via third person report). Mentions were classified as uncertain, when aspects of suicidality were referred to, but did not appear to relate to, risk of the young person being suicidal, for example references to dreams of being dead, or joking about death, or when references to suicidality were about other people (e.g. family members or friends).

Figure 3.2 Overall workflow of the study



Step 2: Manual review of suicidality-related (SR) information and NLP screening tool development.

A randomly extracted subset of 100 patients and their corresponding documents were allocated to a training corpus, and another random selection of 100 patients was allocated to the test corpus. To generate a subset of patients with a reasonable amount of documentation for manual review, the random sample was extracted based on documentation prevalence: each included patients who had at least 7 documents (1st quartile) and at most 50 (3rd quartile), yielding a total of 2,445 (training set) and 2,433 (test set) documents in total. All SR expressions, and labelled each SR-expression as either positive, negated or uncertain were then annotated, according to the rules developed in step 1. However, for this phase, only annotations for SR information (regardless of polarity) were used for analysis.

Step 3: Extension and provisional assessment of the NLP approach for SR screening.

Results from the manual review were used to extend the NLP approach with the addition of new explicit SR expressions. Given the low frequency of the positive or negated SR mentions within the training set, I used the test set to assess precision, recall and F1-score of the tool

detecting any SR content (regardless of polarity). Because the end goal was to address overall suicidality risk behaviour, the approach was evaluated on a document and patient level rather than on the mention level.

Step 4: NLP tool deployed to screen for SR documents

The NLP tool was then deployed to filter out documents without any SR mentions (positive or negative) from the original cohort (excluding the already annotated 200 patients). From 1706 patients (225,577 documents), 890 (52.2%) patients had at least one SR document, resulting in a total of 10,749 documents.

Step 5: Manual review of SR subset for identification of positive (SR-Pos) and negative (SR-Neg) suicidality mentions

Two mental health clinicians under my supervision were randomly assigned 250 (56.2%) patients each from the SR subset. Each clinical annotator was given, for each patient, all documents identified by the NLP tool as containing a SR mention. The annotators were not given the NLP system output, but instead were asked to annotate explicit mentions of suicidality (according the classification rules above) and label these as SR-Pos, SR-Neg or SR-U. The documents were given to the annotators on a per-patient basis, and each patient was reviewed by one annotator. A subset (n=100) of randomly extracted documents was also used to calculate inter-rater agreement [measured with Cohen's kappa (κ) and F1-score] on a document-level.

A majority rule was applied when evaluating document-level agreement: all mention-level annotations in each document were first counted, then, if the number of annotations labelled as positive for suicidality outnumbered or equalled the number of annotations labelled as negated, the document-level label was assigned SR-Pos, otherwise it was designated SR-Neg. To evaluate patient-level performance, priority was given to document-level outcomes: if the patient had at least one document labelled as SR-Pos using the majority rule, the patient-level label was assigned SR-Pos, irrespective of the number of previous or subsequent documents labelled as SR-Neg, i.e. each patient only required a single document to be labelled SR-Pos.

Step 6: Final, comprehensive evaluation: NLP SR-Pos/SR-Neg classification

As a final step, the NLP approach was evaluated with precision, recall and F1-score against the manual annotations of the larger, filtered set of documents/patients with SR-Pos and SR-Neg labels, using the same heuristics for document- and patient-level classification assignments as above. Note that the evaluation is only performed on these two labels, i.e. SR-U annotations are not mapped to SR-Pos or SR-Neg. Thus, a false positive or false negative from the NLP approach could be due to an annotation marked as SR-U. A manual error analysis on cases of disagreements between the NLP tool and human annotation labels was also performed to gain a deeper understanding of the results.

3.3 RESULTS

3.3.1 Distribution of SR annotation within the random selection of test and training set documents

Table 3.1 shows the distribution of SR and NSR documentation and the individual level prevalence amongst the 100 adolescent patients with ASD in the final training set and the 100 patients in the test set. Manual review of both training and test documents revealed that only a small proportion of the corpus contained any SR information: <3% at the document level and around 22% at the patient level, with a similar distribution in the training and test set. Precision, recall, and F1 scores showed high system performance (> 0.8) for both SR and NSR in the test set (table 3.1).

3.3.2 Adaptions to the NLP tool following test and training

The lexical markers of suicidality that were added to the NLP tool included *kill himself/herself/themselves/myself*, *end his/her/their life*, *take his/her/their own life*, *want to die*, *were dead*. Note that the NLP tool relies on lemmatised forms in both target expressions and the document surface forms (i.e. how meaning is expressed by text in the records) in order to achieve a more robust matching. For example, the verb *to want* may appear as *wanted*, *wanting*, *wants*. The base form, 'want' is the *lemma* for the word. Lemmatisation attempts to select the correct lemma depending on the context, so the word "wanting" can be either the base form of a noun or a form of a verb (*to want*) depending on the context; e.g., "*he felt elements of his*

suicide crisis plan were found wanting" or "*He admitted to wanting to end his life*." In this application, lemmatisation enabled *wanting* in the latter sentence to be contextualised as a verb, hence improve the likelihood of a true positive result.

3.3.3 Performance NLP tool on SR test and training set documents

Table 3.2 shows the distribution of negated and positive suicidality-related information (SR-Pos/SR-Neg) using the majority rule criteria in 4,911 pre-screened documents derived from 500 patients. Evaluation of the NLP tool (table 3.2) showed high system performance for SR-Pos with precision, recall, and F1 scores all > 0.83 at a document and patient level. SR-Neg performance measures were lower, especially in recall (0.75 on document level, 0.62 on patient level), but overall good levels of classification were produced (F1 = 0.79 on document level, 0.72 on patient level).

Table 3.1 Confusion matrix: Screening for suicidality (SR) or non-suicidality (NSR), NLP tool compared to human annotation (A).

		NLP (Training)						NLP (Test)					
		Documents			Patients			Documents			Patients		
		NSR	SR	∑	NSR	SR	∑	NSR	SR	∑	NSR	SR	∑
Human Annotation	NSR	2374	10	2384	75	2	77	2356	13	2384	73	5	78
	SR	5	56	61	0	23	23	8	56	64	1	21	22
	∑	2379	66	2445	75	25	100	2365	69	2443	74	26	100
	Precision	0.99	0.85		0.99	0.92		0.99	0.81		0.99	0.81	
	Recall	0.99	0.91		0.97	0.99		0.99	0.88		0.94	0.95	
	F1	0.99	0.88		0.99	0.96		0.99	0.84		0.96	0.88	

Note: NLP; Natural Language Processing

3.3.4 Manual review of NLP and gold-standard discrepancies

A manual error analysis on a random sample of ten documents where the NLP tool classified a document as SR-Pos but the human annotator as SR-Neg was performed to gain a deeper understanding of the reasons behind the lower recall results. The main themes involved:

1) Classification of documents with only one suicide-related mention (annotator SR-Neg count = 1, NLP SR-Pos count = 1) due to missing negation term, e.g. '*Nil* ***suicidal****'* or error in syntactic parsing due to e.g. badly formatted sentences.

2) Cases where the majority heuristic is problematic and the NLP classification of a double negative is erroneous, e.g. one document annotated with SR-Neg = 2, while the NLP output was: SR-Neg = 1, SR-Pos = 3 contained the following: '*XXX denied any recent sleep difficulties, excessive fatigue or guilt, changes in appetite or morbid or* ***suicidal*** *ideation'*, '*The risk of* ***suicide*** *is low, XXX denies* ***suicidal*** *ideation*.'

3) Co-reference in combination with majority heuristics (annotator SR-Neg = 3, SR-Pos = 2, NLP SR-Neg = 1, SR-Pos = 2): '*XXX reported that XXX has had* ***suicidal*** *thoughts in the past but has no current plans on acting on* ***them****$_{co\text{-}reference}$'* (sentence repeated twice in document), '[clinician reporting] *further stated that* ***no*** *evidence of psychosis, self-harming behaviour,* ***suicidal*** *thoughts, sleep or appetite …*'

4) Clinically challenging cases and complex information given in the document. Two examples are:

   (i) Sentences with information reported by external authorities such as the health care team and the school, references to the past, and includes a conclusive statement towards the end of the document: Annotator: SR-U = 2, SR-Neg = 1, NLP tool output: SR-Neg = 2, SR-Pos = 3 '*we could not assess*$_{negated}$ ***suicidal*** *ideation as XXX left the room'*, '*unable to assess*$_{negated}$ ***suicidal*** *ideation'*, '*historically*$_{past}$ *has threatened self harm and disclosed* ***suicidal*** *ideation…'*,

'*concerns from school about **suicidal** ideation'*, '*no **suicidal** ideation expressed'*.

(ii) No clear opinion expressed by the patient or the clinicians : Annotator: SR-U = 2, SR-Neg = 1, NLP tool output: SR-Pos = 2: 'I tried to assess XXX's **suicidal** risk - XXX does not know if XXX wants to **kill XXXself**', 'XXX does not have any specific plan'

Table 3.2 Classification of positive and negative suicidality, document- and patient level assessments.

		NLP (Test)					
		Documents			Patients		
		SR-Neg	SR-Pos	∑	SR-Neg	SR-Pos	∑
Annotator	SR-Neg	1379	463	1842	81	50	131
	SR-Pos	273	2796	3069	14	355	369
	∑	1652	3259	4911	95	405	500
	Precision	0.83	0.86		0.85	0.87	
	Recall	0.75	0.91		0.62	0.96	
	F1	0.79	0.88		0.72	0.92	

Note: SR-Neg; Suicidality-related mention is negated. SR-Pos; Suicidality-related mention is positive.

In total, 100 random documents were double annotated (table 3.3). A document-level assessment using the majority rule yielded an average Cohen's κ of 0.83, F1-scores for SR-Neg and SR-Pos document-level assessment were 0.89 and 0.94 respectively, indicating high agreement.

Table 3.3. Confusion Matrix: Inter-Rater Agreement on document level. SR-Neg = Suicidality-related (SR) mention is negated (Neg), SR-Pos = Suicidality-related mention is positive (Pos).

		Annotator 1		
		SR-Neg	SR-Pos	∑
Annotator 2	SR-Neg	32	2	34
	SR-Pos	6	60	66
	∑	38	62	100

3.4 DISCUSSION

This is the first study to demonstrate that an NLP tool can be used to accurately capture a clinical construct as complex as suicidality within health records of young people with ASD. The NLP tool identified suicidality-related (SR) mentions with high degrees of precision (0.81) and recall (0.84) from clinical free text documents held within EHRs. This NLP application provides powerful opportunities for surveillance work in adolescent ASD and in other clinical samples, with the potential to improve risk prediction for major adverse events, such as suicide attempts.

The development of this high-performance NLP tool was achieved in several steps. First, owing to the potentially distinctive characteristics of the ASD clinical population, and their specialist mental health service provision, I began by building a suicidality terminology from a detailed note review of over 2000 random sets of clinical entries in 100 children with ASD, combined with expert clinical consensus. Because of the limited literature on suicidal terminology in ASD, I used a randomly extracted training and test set from all potential ASD EHR source data, rather than an enriched set filtered by restricted terms (e.g. "suicid*" [159] or ICD coding classifications).[161] The rationale for this was to reduce selection bias and loss of sensitivity through the use of training and test data derived using restricted terms or coding classifications. Random selection from the whole potential corpus also provided me with a better understanding of the overall distribution of suicidality-related information in documents, and allowed us to refine and advise on additional terminology. During the training phases, it became clear that there was a low frequency of SR terms (less than 3% of all documents). A much larger corpus was then required to conduct an adequate test of the NLP tool's classification performance in discerning positive and negated SR mentions within the documents.

The abstraction of mention-level annotations and NLP system predictions to document- and patient-level assessments using simple heuristics (majority rule for document level and SR-Pos priority on patient level) showed that promising results can be obtained even though the NLP tool relies only on a relatively small number of suicidality-related and negation terms. This

finding also shows that even though suicidal behaviours are documented with a variety of expressions (e.g. '*took an excessive amount of pills*', '*threw him/herself in front of a train*'), indicative terms (mainly *suicide* in different forms) are typically also used at some point in the documentation, and will thus be eventually detected automatically.

3.4.1 Strengths

A strength of this study is that I have not assumed that clinical terms used in more typically developing children or adults generalize to ASD populations. In practice, assessing suicidality in adolescents with ASD often requires a different approach to other patient groups, which in my clinical experience was likely to be reflected in the clinical notes. Young people with ASD presenting to mental health services commonly have severe difficulties with interpersonal interactions, making for a more complex clinical assessment.[162] Clinicians are likely to deliberate within the clinical notes on whether potential behaviours are driven by suicidal ideation, potentially creating more false positive results. They may have a greater reliance on third person report – i.e. caregivers voicing concerns regarding the young person's suicidality rather than direct accounts from the young person. Also, where a first-person account is provided, clinicians will often write verbatim statements (e.g. He told me "*I just want to end it*", and he "*went to the car park to get it done*"), providing more atypical clinical terminology for describing suicidality, and increasing the chance of NLP misclassification.

In addition, young people with ASD may not present with suicidality as a principle complaint, but through a behavioural change such as school refusal, with suicidal behaviour emerging through later clinician screening. This may change the emphasis and position within the patient's clinical record relative to other populations where suicidal behaviour is the principle trigger during the first presentation to services. Testing these clinical assumptions empirically using a non-ASD control sample was beyond the scope of the current study, however future work is underway to examine the variability of the NLP tool's accuracy across non-ASD child populations seen in mental health services. NLP applications are commonly validated using randomly extracted documents from EHRs covering a broad range of clinical contexts, seldom rarer clinical populations, such as young people with ASD. As mental health assessment and management needs to be tailored to the developmental needs of the young people in clinic, so should the validation of NLP data extraction tools.

3.4.2 Limitations

The motivation for applying a majority rule on document level assessments was based on the finding that the main source for false positive errors in the negation detection approach stemmed from cases of question forms (e.g. '*I asked him if he feels suicidal*'), references to the past, etc. Applying this rule was a way of smoothing this error rate. However, the error analysis showed that this approach might be a limitation. In future studies, I aim to compare results with NLP approaches such as ConText[163] where variables relating to the past ('historicity') and subject ('experiencer') are encoded with target terms. I also aim to experiment with other abstraction heuristics, e.g. instead of majority rule, applying a priority hierarchy. In keeping with prior work, another alternative could be to define the annotation task on a document level.[157] Longer term, I aim to compare the predictive validity of different heuristics within the NLP tool, and across other NLP approaches, for later adverse outcomes (i.e. significant suicide attempts or death by suicide), and seek external validity through replication in other EHR systems. Without these further steps, it is difficult to assess the potential clinical impact of differences in precision or recall across NLP tools.

The clinical annotators I supervised expressed that it was sometimes challenging to assess suicidality risk based on one document at a time; single documents did not provide sufficient context in all cases. At the same time, given the rare prevalence of suicide-related content in all patient documents, defining a patient-level annotation task using this type of abundant clinical documentation would be very time-consuming. I plan to explore different ways of addressing this issue, one being a nested case-control study design similar to the one presented in Metzger et al.[95]

3.4.3 Conclusion

The suicidality outcome data provided by this NLP extraction tool permits analyses of the complex interplay of ASD-specific traits on factors contributing to the onset and recurrence of suicidality. ASD specific mental health services are becoming increasingly available for child and adolescent populations in high-income countries. Although there is more work to be done before clinical application, we believe the NLP tool described provides a step forward in enhancing suicidality surveillance, risk prediction and treatment selection for children with ASD.

CHAPTER 4. THE ASSOCIATION BETWEEN CO-MORBID AUTISM SPECTRUM DISORDERS AND ANTIPSYCHOTIC TREATMENT FAILURE IN EARLY-ONSET PSYCHOSIS: A HISTORICAL COHORT STUDY USING ELECTRONIC HEALTH RECORDS.

The contents of this chapter have contributed to the following:

Publication in a peer-reviewed journal

Downs J, Lechler S, Dean H, Sears N, Patel R, Shetty H, Simonoff E, Hotopf M, Ford T, Diaz-Caneja MD, Arango C, McCabe JH, Hayes RD, Pina-Camacho L. The association between co-morbid autism spectrum disorders and antipsychotic treatment failure in early-onset psychosis: a historical cohort study using electronic health records. *Journal of Clinical Psychiatry* (in press)

4.1 SUMMARY

Background: In a sample of children and adolescents with first-episode psychosis, I investigated whether multiple treatment failure (MTF, defined as the initiation of a third trial of novel antipsychotic due to non-adherence, adverse effects or insufficient response) was associated with co-morbid autism spectrum disorders.

Methods: Data were from the electronic health records of 638 children (51% male) with first-episode psychosis, aged between 10 and 17, referred to mental health services in South London, UK, using the CRIS system. The effect of autism spectrum disorder comorbidity on the development of MTF over a 5-year period was modelled using Cox regression.

Results: There were 124 cases of MTF prior to the age of 18 (19.3% of the sample). Co-morbid autism spectrum disorders were significantly associated with MTF (adjusted hazard ratio aH.R 1.99, 95% CI 1.19–3.31; p=0.008) after controlling for a range of potential confounders. Other factors significantly associated with MTF included older age at first presentation, Black ethnicity, and frequency of clinical contact. No significant association between other co-morbid neurodevelopmental disorders (hyperkinetic disorder or intellectual disability) and MTF was found.

Conclusions: Among children with first-episode psychosis, those with co-morbid autism spectrum disorders at first presentation are less likely to have a beneficial response to antipsychotics.

4.2 INTRODUCTION

Nearly a fifth of individuals diagnosed with a psychotic disorder experience their first episode under 18 years of age.[164] Relative to adults with first-episode psychosis, children appear to have a significantly worse symptomatic and functional recovery,[165] hence early-onset psychosis (EOP) may represent a more severe form of the disorder. Comparisons between first-episode psychosis in adult and child cohorts show children have poorer premorbid functioning or adjustment,[164,166] greater cognitive deficits,[167] more primary negative symptoms at first presentation[165] and - albeit less consistently replicated - longer durations of untreated psychosis.[168] It is these factors that appear to be most consistent predictors of poor clinical and functional improvement in EOP samples at follow-up.[169]

Premorbid difficulty is a broad construct, often retrospectively ascertained, which encompasses childhood history of developmental milestone delays, poor sociability, poor peer relationships, limited scholastic performance, problems with adaptation to school, and socio-sexual development.[170,171] Specific neurodevelopmental conditions, such as autism spectrum disorders (ASD), which, by definition, represent the extreme manifestations of poor premorbid difficulties,[172] elevate the risk of developing psychosis.[173–175] Whilst premorbid difficulties have been associated with poor outcomes in both early-onset [176,177] and adult-onset psychosis,[178,179] the mechanism of how it affects psychosis prognosis is unclear. One possibility is that premorbid function is associated with lower responsiveness to antipsychotic treatment, with recent evidence showing poorer premorbid function is a predictor of adult treatment resistant schizophrenia.[180]

The effect of autism spectrum disorder on treatment effectiveness has not been examined in early onset psychosis samples. This represents an important gap in the evidence, as work in non-psychotic conditions suggests that psycho-pharmacological effectiveness is lower in populations with co-existing ASD.[181] Furthermore children and adolescents with mixed ASD-psychotic profiles are not uncommon in clinical practice.[175] Recent studies show that ASD may be present in 30-50% of children diagnosed with severe psychotic disorders.[182]

I conducted a longitudinal study which aimed to investigate whether co-morbid ASD was associated with a pragmatic measure of poor antipsychotic treatment response in a large

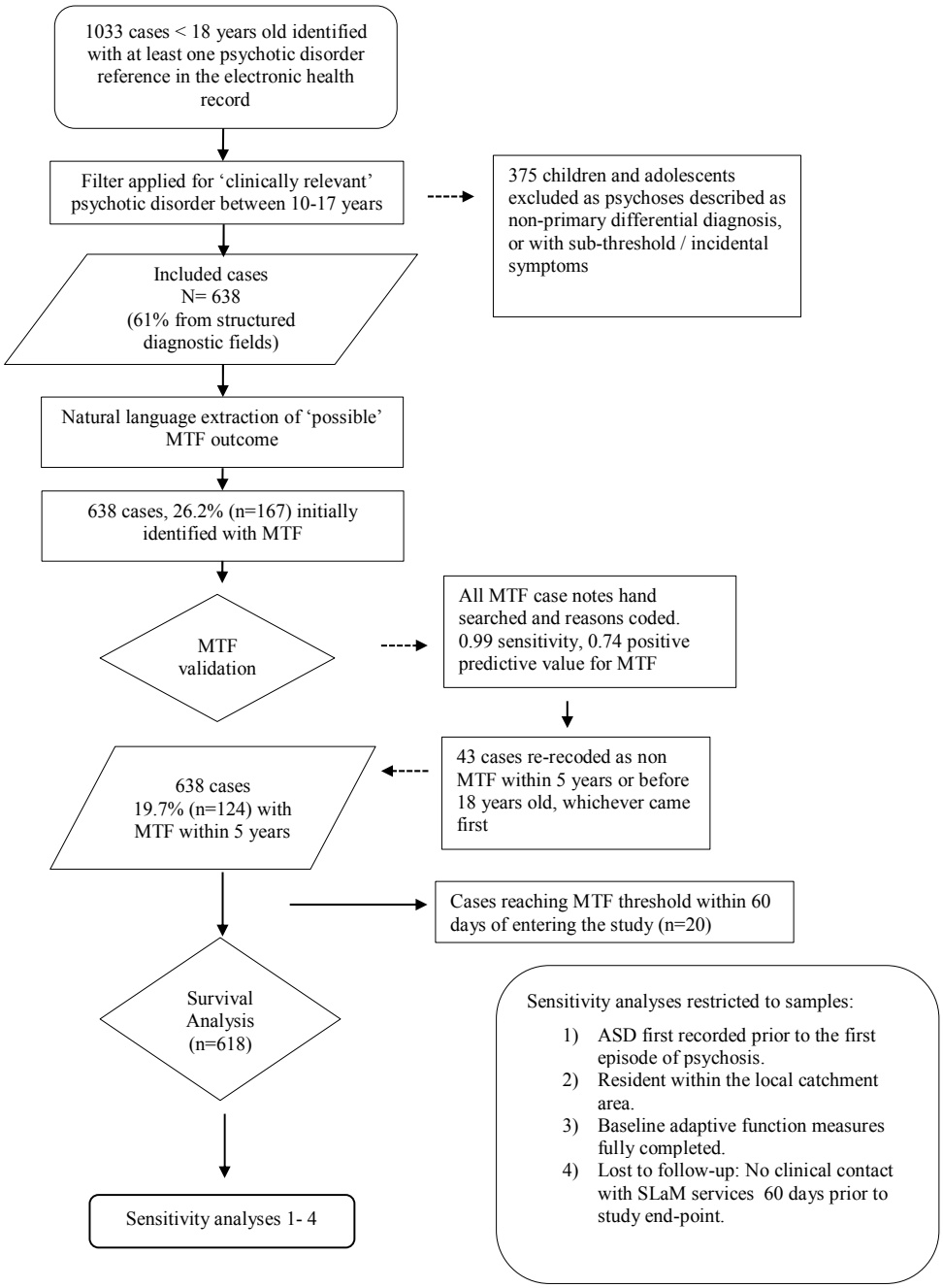
historical clinical cohort of children and adolescents with first-episode psychosis. I predicted that patients with co-morbid ASD would be more likely to experience treatment failure. I also expected that this association would remain after taking account of potential confounders, including psychotic disorder category, and additional markers of premorbid neurodevelopmental difficulties such as co-occurring hyperkinetic disorder and intellectual disability.

4.3 METHODS

4.3.1 Study Setting

This study used data extracted from the electronic mental health records of an open cohort of children and adolescents referred to SLaM CAMHS, with a first episode of any psychotic disorder between 1st January 2008 and 1st November 2014. CAMHS comprised of inpatient, outpatients and early intervention for psychosis services. Over this period, SLaM provided all aspects of specialist mental healthcare to a catchment population of approximately 250,000 children resident within four London boroughs (Lambeth, Southwark, Lewisham, Croydon). In addition to the district services, SLaM provided specialist inpatient and outpatient mental health assessment and treatment services for young people from outside the local district. Each borough had a dedicated multidisciplinary service for children, which accepted referrals for school age children (4–18 years; exceptionally cases are accepted below this age) with suspected or previously confirmed neurodevelopmental disorders, displaying emotional or behavioural difficulties. Children were referred from primary care, child health, and educational and social care services, and typically underwent a multidisciplinary assessment by CAMHS clinicians.

Figure 4.1 Flow chart of study inclusion and analysis



4.3.2 Study sample

The sample data were extracted using the CRIS system, which provided access to a de-identified record database containing the electronic mental health records over 35,000 child and adolescent cases (see chapter 2 for more details).[129]

Figure 4.1 shows the flowchart for inclusion in the study. All cases who had presented to SLaM services aged between 10-17 years, were screened for ICD-10 diagnoses within clinician-

recorded structured or unstructured free text fields. Those with structured data recorded were included if they had at least one psychosis diagnosis (ICD-10 codes F20-F29, F30-31, F32.3, F33.3, F1x.5). Missing structured diagnostic data was supplemented by GATE (Generalized Architecture for Text Engineering), a natural language processing tool which codes 'free text' diagnostic data.[130] GATE extracted all CAMHS records with any free text diagnosis of "schizophrenia, schizoaffective disorder, bipolar disorder, depression with psychosis symptoms, acute and transient psychosis, delusional disorder, induced delusional disorder, drug-induced psychosis and psychoses not otherwise specified (NOS)." These were filtered for any clinician-recorded mention of antipsychotic treatment after the psychosis diagnosis. This process reduced the inclusion of children with non-psychotic indications for antipsychotic use, psychoses as differential diagnoses, and sub-threshold/incidental psychotic symptoms. Out of the 1033 cases identified with at least one psychotic disorder recorded, only 638 individuals with a 'clinically relevant psychotic disorder' were included (see figure 4.1). The earliest recorded psychosis diagnosis was coded as the first diagnosis. A hand-searched review of a random sample of 100 records revealed this identification process provided a 0.98 positive predictive value (PPV) for psychotic disorder diagnosis.

For each participant, the study entry date was the accepted referral date to CAMHS for their first-episode psychosis. Baseline exposure data (i.e. clinical and socio-demographic data) were drawn from all notes entered within 60 days of study entry. The follow-up period ran from 60 days after their accepted referral date to the date of their 18th birthday, date of death, or the end of the 5-year observation period (whichever came first). Frequency of clinical contact during the follow-up period was determined through the days each person had received face-to-face contact as recorded in structured fields. Multiple events on a single day were counted as one day of clinical contact, whilst clinical contact with outpatient services during an inpatient admission was not counted.

4.3.3 Measurements

Outcome: multiple antipsychotic treatment failure

In contrast to standard definitions in adults for treatment response, no established minimum antipsychotic therapeutic dose thresholds or treatment periods existed for children and adolescents with psychosis,[183] similarly no standard criteria for poor antipsychotic response or refractory disorder [184] appeared suitable to a retrospective cohort study of EOP using electronic health records.[185,186] Therefore I created a proxy, based on the antipsychotic effectiveness literature,[187–189] which I termed 'multiple treatment failure' (MTF). I defined MTF as the initiation of a third trial of a novel antipsychotic due to insufficient response, intolerable adverse effects or non-adherence to prior antipsychotic treatment. A previously validated GATE application was used to identify novel regular antipsychotic prescription trials as a replacement or adjunctive treatment to the previous trial, this excluded antipsychotic medication prescribed on an 'as required basis' [132,160] or switching preparations – e.g. oral to depot administration. The date of MTF was determined when a third novel antipsychotic medication was started within a 5-year follow-up period.

I performed further validation alongside other clinical raters under my supervision. The raters, blinded to MTF status, hand-searched 100 cases from the sample each, which included all 167 individuals (55-56 per rater) where MTF was initially identified, and a random selection of non-MTF individuals (44-45 per rater). The GATE identification process provided >0.99 sensitivity for MTF (i.e. no false negatives) and 0.74 PPV. False positives largely occurred where antipsychotic medications were used for non-psychotic indications. These cases (n=43 subjects, 6.7% of the total sample) were subsequently recoded as non-MTF. Raters also manually coded the reasons for treatment failure for each novel antipsychotic trial in the MTF group, and coded the predominant reason. Consistent with previous literature,[190] reasons were defined as insufficient response, intolerable adverse effects, non-adherence, 'other' and 'reasons not identified'. For 15 randomly selected cases, first and second treatment failure reasons were coded by two raters independently. Percentage agreement ranged from 0.67 to 0.87. Kappa coefficients indicated agreement from moderate, for adverse effects at first treatment failure ($\kappa = 0.33$), to substantial for insufficient response at second treatment failure ($\kappa = 0.71$). Within the MTF group, those cases identified as having the same reason for antipsychotic discontinuation/switch at first and second trials were grouped into four MTF

'persistent reason' groups. A 'variability in reasons' subgroup (i.e. when reasons were different at each antipsychotic trial) was also created.

Extraction of ASD comorbidity data

Clinician recorded ASD comorbidity (ICD-10 F84.0, F84.1, F84.5-9) was extracted from the clinical record at any time point during the observation period, using free text and structured fields.[160] Compared with expert consensus, prior work has established a high specificity for ASD diagnoses by clinicians working at a district level.[125] Patients were included in the ASD group if they fulfilled ICD-10 criteria for Pervasive Developmental Disorder after direct clinical observation and taking a full psychiatric and developmental history from at least one informant, typically the mother. The Autism Diagnostic Observation Schedule (ADOS)[191] was administered by experienced ADOS trained clinicians when the diagnosis was not clear (52 cases). The final diagnosis was based on best clinical judgment considering all the available information,[192] by NHS clinicians certified to administer the Autism Diagnostic Interview[193] and research-certified to administer the ADOS. Additional validation of ASD diagnosis data extraction was carried out by a hand search of the 100 randomly selected cases. The data extraction methodology was found to have a high sensitivity (0.82) and PPV (0.86).

Extraction of Covariates: clinical and other demographic data

A number of demographic and clinical variables were extracted at baseline (i.e. within 60 days of study entry). Demographic variables included gender, age at referral for first-episode psychosis, ethnicity (categories defined by the UK Office for National Statistics), and index of neighbourhood deprivation for the main caregiver residence.[194]

The first clinically recorded ICD-10 psychosis diagnoses were grouped into schizophrenia, schizoaffective disorder, bipolar disorder, depression with psychosis symptoms specified, drug-induced psychosis, and other psychoses. Other neurodevelopmental disorder comorbidities extracted included hyperkinetic disorders (ICD-10 F90) and intellectual disability (ICD-10 F70-9). Inpatient admission status and adaptive function - ascertained using the Children's Global Assessment Scale (CGAS)[133] within 60 days of study entry - were also extracted.

4.3.4 Analysis

To compare demographic, clinical characteristics of individuals and MTF outcomes, with and without co-morbid ASD, crude analyses were conducted using chi squared for categorical variables, and Student's independent t-test for continuous variables. To examine the prospective association between baseline demographic, clinical exposures and MTF outcome, I excluded children who had MTF within the 60-day baseline period (*n*=20). After checking proportional hazards assumptions, I used a Cox regression to model the association between ASD comorbidity and MTF. The first model examined the crude effect of ASD alone on MTF. Subsequent models were constructed adding potential socio-demographic and clinical confounders. Fully-adjusted survival hazards and separate survival curves were plotted to compare the risk of MTF between children with and without ASD comorbidity.

To account for the potential effect of diagnostic re-classification of psychosis to ASD explaining any association between ASD and MTF, I conducted a sensitivity analyses by removing the sample of children with ASD first recorded 30 days after the first psychosis diagnosis date (n=48). Three additional sensitivity analyses were conducted: i) to restrict the analyses to children with complete adaptive function measures (CGAS) at first presentation (*n*=394), see figure 4.1, as this could be a potential confounder for any ASD-MTF association; ii) to test whether being resident within the local catchment area (as opposed to children referred from outside the 4 local districts) had an effect on the association between ASD and MTF, as families residing outside the local catchment area can receive additional non-SLAM mental health service not captured within SLaM health record system; iii) to test whether being potentially lost to follow-up by SLaM services (n=295, defined as no clinical contact within 60 days of the study end point) had an effect on the association between ASD and MTF.

4.4 RESULTS

4.4.1 Demographics and clinical characteristics of the sample

I identified 638 young people (329 male) aged between 10 and 17 years with a clinically relevant psychosis diagnosis (figure 4.1). The average follow-up period was 1.79 years (SD 1.4, range 0.1-5). Out of those, 124 (19.4%) developed MTF during the follow-up period, at a mean age of 16.3 years (SD 1.4). Table 4.1 provides further information on the socio-demographic and clinical characteristics of the total sample and the subsample eventually developing MTF.

Table 4.1 Demographic and clinical characteristics of young people with first-episode psychosis (n=638)

Sample characteristics	Total Sample (n=638, %)	MTF (n=124, %)
Gender, n (%)		
Male	329 (51.1)	59 (47.2)
Female	309 (49.9)	65 (53.8)
Mean age at referral (SD)	15.6 (1.9)	15.4 (1.6)
Mean age of reaching MTF (SD)	---	16.3 (1.4)
Mean years of follow-up (SD)	1.79 (1.4)	2.1 (1.2)
Mean clinical contact days (SD)	93 (112)	205 (147)
Ethnicity, n (%)		
White British	260 (40.8)	44 (35.5)
White Other	37 (5.8)	7 (5.6)
Black	209 (32.8)	51 (41.1)
Asian	39 (6.1)	7 (5.6)
Mixed	74 (11.6)	15 (12.2)
Not Stated	19 (2.9)	0%
Neighbourhood Characteristics, n (%) [a]		
1st (Least Deprived)	165 (26.6)	39 (32.5)
2nd	152 (24.6)	28 (23.3)
3rd	151 (24.4)	25 (20.8)
4th (Most Deprived)	151 (24.4)	28 (23.4)
First ICD-10 psychosis diagnosis, n (%)		
Schizophrenia	365 (57.1)	63 (50.8)
Bipolar Disorder	42 (6.6)	9 (7.3)
Schizoaffective	17 (2.7)	10 (8.1)
Psychotic Depression	69 (10.8)	14 (11.2)
Drug induced psychosis	39 (6.1)	6 (4.8)
Other Psychoses	106 (16.6)	22 (17.8)
Co-morbid neurodevelopmental disorders, n (%)		
Autism Spectrum Disorder	114 (17.9)	33 (26.6)
Hyperkinetic Disorder	40 (6.3)	<5%
Intellectual Disability	65 (10.2)	15 (12.1)
Baseline function		
Admission at first presentation n (%)	260 (40.8)	72 (58.1)
Children's Global Assessment mean score (SD)[b]	38.3 (15.9)	35.1(16.0)

Note: Standard Deviation (SD); Missing cases = [a] 19, [b] 216

Table 4.2 Demographic and clinical characteristics of first-episode psychosis in young people with and without co-morbid autism spectrum disorder (n=638)

Sample characteristics	Autism Spectrum Disorder		
	No (n=524)	Yes (n=114)	*P* Value*
Multiple treatment failure (MTF), n (%)	91 (17.4)	33 (29.0)	0.005
Mean age at referral (SD)	15.8 (1.7)	14.5 (1.8)	<0.001
Mean age of reaching MTF (SD)	16.4 (1.3)	15.9 (1.4)	0.04
Mean years of follow-up years (SD)	1.61 (1.3)	2.6 (1.4)	<0.001
Mean clinical contact days (MTF)	90 (108)	109 (128)	0.04
Male gender, n (%)	254 (48.5)	75 (65.9)	0.001
Ethnicity, n (%)			
White British	209 (39.8)	51 (44.7)	0.34
White Other	34 (6.5)	3 (2.6)	
Black	174 (33.2)	35 (30.7)	
Asian	32 (6.1)	7 (6.1)	
Mixed	62(11.8)	12 (10.5)	
Not Stated	13 (2.5)	6 (5.3)	
Neighbourhood Characteristics, n (%) [a]			
1st (Least Deprived)	130 (25.6)	35 (31.5)	0.55
2nd	124 (24.4)	28 (25.2)	
3rd	129 (25.4)	22 (19.8)	
4th (Most Deprived)	125 (24.6)	26 (23.4)	
First ICD-10 psychosis diagnosis, n (%)			
Schizophrenia	316 (60.3)	49 (43)	<0.001
Bipolar Disorder	34 (6.5)	8 (7.0)	
Schizoaffective	14 (2.7)	3 (2.6)	
Psychotic Depression	55 (10.5)	14 (12.3)	
Drug induced psychosis	38 (7.3)	1 (0.9)	
Other Psychoses	67 (12.8)	39 (34.2)	
Baseline function			
Admission at first presentation, n (%)	228 (43.4)	32 (28.1)	0.002
Children's Global Assessment Scale (CGAS)[b] mean (SD)	38.4 (16.1)	37.4 (15.0)	0.32
Other neurodevelopmental disorders, n (%)			
Hyperkinetic Disorder	22 (4.2)	18 (15.8)	<0.001
Intellectual Disability	35 (6.9)	30 (26.3)	<0.001

*χ^2 *tests* for categorical variables and Student's independent t-test for continuous variables; Missing cases = [a] 19, [b] 216

4.4.2 Characteristics of the sample by ASD status

Characteristics of co-morbid ASD (*n*=114) vs non-ASD (*n*=524) subsamples are provided in table 4.2 Twenty-nine percent of the sample with co-morbid ASD developed MTF compared to the 17% of the non-ASD sample (p <0.01), and reached MTF at an earlier age (p<0.05). Details on the antipsychotic treatment pathways for the 124 children who developed MTF are provided in table 4.3 and in table 4.4. The largest proportion (47%) switched their first antipsychotic due to intolerable side effects, whilst 21% showed insufficient response. After the second antipsychotic trial, nearly one third of MTF children had an insufficient response (table 4.3).

Table 4.3 Reasons for switching at first and second trial of antipsychotic treatment in young people with first-episode psychosis who develop multiple treatment failure (MTF, n=124).

Reasons for changing antipsychotic treatment	Individuals with Multiple Treatment Failure	
	1st to 2nd antipsychotic treatment	2nd to 3rd antipsychotic treatment
Insufficient response n (%)	26 (21.1)	39 (31.7)
Intolerable adverse effects n (%)	55 (44.7)	39 (31.7)
Non-adherence n (%)	18 (14.6)	19 (15.5)
Other reason / No reason ascertained n (%)	25 (20.2)	27 (21.8)
Median duration in days (25-75th centile) before change to novel antipsychotic treatment	76 (20-272)	121 (41-314)

4.4.3 Pathways to antipsychotic treatment failure

Table 4.4 provides information regarding treatment failure pathways. The three main patterns of discontinuation were the combination of insufficient response and adverse events over time (*n*=28, 32.6%) persistent adverse effect (*n*=18, 21%) and persistent insufficient response (*n*=13, 15.1%) trajectories, with significant differences in the reasons for MTF between ASD and non-ASD groups (p =0.05). Children with ASD showed higher rates of the 'persistent insufficient response' or the 'insufficient response-adverse effect' trajectory but lower adherence-related reasons relative to those without ASD (table 4.4).

Table 4.4 Reasons for multiple treatment failure (MTF) in young people with first-episode psychosis, with and without co-morbid autism spectrum disorder

Reasons for MTF[a]	n (%) of individuals [b]	
	No Autism Spectrum Disorder (n=65)	Autism Spectrum Disorder (n=21)
Persistent insufficient response	7 (10.8)	6 (28.6)
Persistent adverse effects	15 (23.1)	4 (19.1)
Persistent non-adherence	5 (7.7)	0 (0)
Variability in reasons		
• Insufficient response and adverse effects	18 (27.7)	10 (47.6)
• Insufficient response and adherence	6 (9.2)	0 (0)
• Adverse effects and adherence	14 (21.5)	1 (4.8)

[a] Comparison in reasons for MTF between No Autism Spectrum Disorder (no ASD) and ASD groups; χ^2 =11.1, df =5, p=0.05

[b] In all cells, % refers to percentages (within columns) of individuals for whom information on main reason of discontinuation was available. Excluded due to no reason ' or 'other reason' ascertained were: No ASD group n=26 (28%); ASD group n=12 (36%)

4.4.4 ASD and the association with MTF

Cox regression models are displayed in table 4.5, and graphically represented in figure 4.2. Comorbid ASD was associated with an increased risk of reaching MTF over the follow-up period (adjusted hazard ratio, (aH.R) 1.99, 95% CI 1.19–3.31; *p*=0.008). This was after adjusting for potential confounders including socio-demographic factors, co-morbid hyperkinetic disorder or intellectual disability and, as a marker of psychosis severity, admission status at presentation and clinical contact over the follow-up. Age at first referral, Black ethnicity, and frequency of clinical contact over the follow-up period were also positively associated with MTF (see table 4.5).

4.4.5 Sensitivity Analysis

From the sensitivity analyses conducted, I found no change in the direction of the effect of ASD on MTF, although the restriction in sample size meant loss of statistical power : i) the subsample of children with ASD diagnosis recorded prior to their psychosis diagnosis (excluding 48 children with co-morbid ASD) aH.R 1.48, 95% CI 0.81–2.73; p=0.2; ii) children with complete CGAS information (n=394) aH.R 1.98, 95% CI 1.06–3.67; p=0.03; iii) children resident exclusively within the local catchment area (n=329) aH.R 1.51, 95% CI 0.69–3.28; p=0.30; iv) children with no clinical contact recorded within 60 days of the study end date (n=295) aH.R 2.71, 95% CI 1.14–6.39; p=0.02.

Figure 4.2: Probability of treatment effectiveness (non-multiple treatment failure) after first-episode psychosis, comparing children with and without autism spectrum disorder (adjusted for all table 4.5 variables)

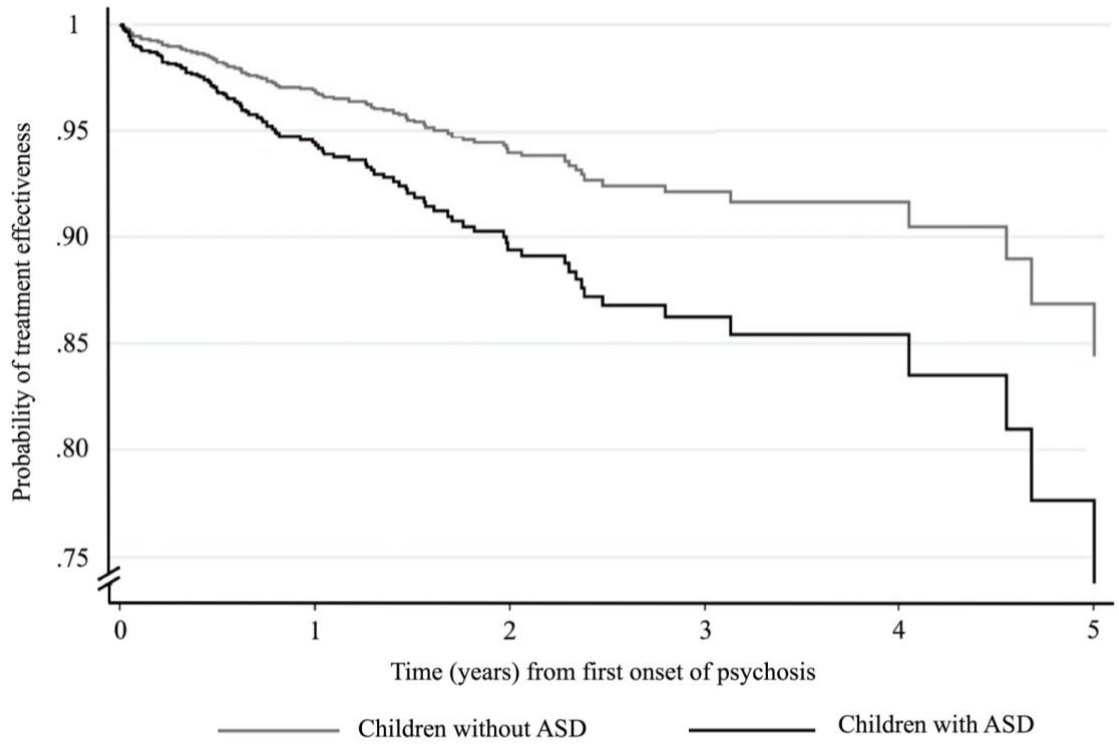


Table 4.5. Multivariable cox regression analysis of the association between autism spectrum disorder and multiple treatment failure in children and adolescents with first-episode psychosis (n=618)

Multiple Treatment Failure	Crude H.R. (95% CI)	*P*	Adjusted for Socio-demographic factors H.R. (95% CI)	*P*	Fully adjusted Model H.R. (95% CI)	*P*
Autism Spectrum Disorder	1.24 (0.80 – 1.90)	0.33	1.52 (0.95 – 2.42)	0.08	1.99 (1.19 – 3.31)	0.008
Female (vs male)			1.18 (0.78 – 1.77)	0.43	1.26 (0.82 – 1.92)	0.29
Age at referral			1.31 (1.31 – 1.52)	<0.001	1.39 (1.19-1.64)	0.001
Ethnicity						
White British			Reference		Reference	
White Other			0.67 (0.21 – 2.11)	0.52	0.92 (0.28 – 3.09)	0.90
Black			2.03 (1.28 – 3.22	0.003	1.73 (1.04 – 2.86)	0.03
Asian			1.20 (0.50 – 2.86)	0.68	1.24 (0.51 – 3.07)	0.63
Mixed			1.50 (0.79 – 2.83)	0.22	1.54 (0.79 – 3.03)	0.20
Not Stated [a]			n/a		n/a	
Neighbourhood Characteristics						
1st (Least Deprived)			Reference		Reference	
2nd			0.64 (0.37 – 1.09)	0.11	0.67 (0.37 – 1.19)	0.18
3rd			0.56 (0.32 – 0.98)	0.04	0.70 (0.38 – 1.28)	0.25
4th (Most Deprived)			0.57(0.32 – 0.99)	0.05	0.72 (0.39 – 1.32)	0.30
First ICD-10 psychosis diagnosis						
Schizophrenia					Reference	
Bipolar Disorder					1.29 (0.61 – 2.73)	0.50
Schizoaffective					1.57 (0.56 – 4.35)	0.38
Psychotic Depression					1.27 (0.67 – 2.39)	0.46
Drug induced psychosis					1.34 (0.47 – 3.81)	0.99
Other Psychoses					0.85 (0.49 – 1.47)	0.55
Other neurodevelopmental disorders (hyperkinetic disorder and/ or intellectual disability)					0.70 (0.38 – 1.27)	0.24
Admitted at first presentation					1.18 (0.78 – 1.81)	0.45
Mean clinical contact days					1.006 (1.004 – 1.07)	<0.001

[a] Variable dropped due to 0 values in cell

4.5 DISCUSSION

This is the first longitudinal study to examine the association between co-morbid ASD and poor antipsychotic treatment outcomes in children with first-episode psychosis. Using electronic health record data from community and inpatient CAMH services, I found that 19% of children developed MTF before the age of 18. I found that ASD co-morbidity was associated with a 2-fold increased risk of MTF, after adjustment for potential sociodemographic and clinical confounders including gender, ethnicity, age at first referral, psychosis subcategory, and illness severity. Among children with MTF, most cases did not show a consistent mechanism of discontinuation over time but, of note, 28% of those with co-morbid ASD compared to 11% of non-ASD children, had a persistently insufficient response to antipsychotics. These findings suggest that the effect of developmental delays and poor premorbid adjustment on antipsychotic treatment failure found in adult studies of first-episode psychosis,[195,196] are applicable to children with early-onset psychosis.

The study findings may be explained by specific neurobiological profiles related to psychosis-ASD comorbidity. Certainly, pharmacological treatments for non-psychotic disorders in ASD appear to have reduced effectiveness.[181] For example, children with ASD tend to respond less favourably to methylphenidate or to antidepressants, and experience adverse effects to these agents more often, and with greater severity, than their peers without ASD.[181,197] ASD-psychosis subgroups may have a reduced dopamine synthesis capacity and diminished response to dopamine receptor blocking antipsychotics.[198,199] These theoretical mechanisms cannot be explored within the data available in this study, but the findings support further investigation into interventions that target alternative non-dopaminergic pathways in children with ASD-psychosis co-morbidity.

I found other predictive factors that were significantly associated with MTF including Black ethnicity, older age at referral (a proxy for age at first episode), and frequency of clinical contact. Children of Black ethnicity were twice as likely as white British, to develop MTF, which is consistent with a number of studies in adults.[200,201] Clinical contact with services was positively associated with a risk for MTF. This is in keeping with other research, in early-onset psychosis samples, which indicates symptom severity and increased service use are associated with a more complicated illness course.[169] Male gender was not associated with an increased risk of MTF, which suggests that it is not a prognostic marker for treatment effectiveness,

although it is a risk factor for psychosis in adolescence. These results accord with a number of studies examining demographic predictors for poor social functioning and treatment resistance, in early-onset[186,202] and adult cohorts.[203]The study findings suggest that most young people with early-onset psychosis do not develop treatment failure via a consistent mechanism of discontinuation. Nearly 60% of the MTF group had different reasons for the discontinuation of each trial of novel antipsychotic. In cross-section, I found similar patterns of discontinuation to other early-onset studies. The Treatment of Early-Onset Schizophrenia Spectrum Study (TEOSS) found 39.2% of the discontinuers experienced an insufficient response, and 36% reported adverse effects.[204] Similarly, I found nearly 32% of children with MTF had switched from the first antipsychotic trial due to intolerable adverse effects, and 32% showed insufficient response.

4.5.1 Strengths

This study has a number of strengths. I studied one of the largest child and adolescent samples presenting with their first-episode psychosis, which permitted us sufficient power and precision to estimate the strength of the association between ASD and MTF, whilst taking account of a number of potential confounders. It was a first-episode sample, hence participants shared a common starting point in their illness course, which reduced the confounding effects of illness duration and unknown treatment exposures typically found in other early- and adult-onset onset schizophrenia cohort studies. Importantly, the findings can be readily generalized to clinical practice. The sample included the whole clinical population of four south London boroughs that were accessing 'real world' inpatient and outpatient CAMH services.

4.5.2 Limitations

Some limitations should be considered when interpreting the results of this study. As with all health record databases, there is some risk that not all clinical details are available for participants throughout the study duration. However, I would expect the data to be representative of children with psychoses living in urban and suburban areas since SLaM is a near-monopoly provider of specialist mental healthcare for its geographic catchment. I drew on complete electronic clinical records for over 600 cases, providing the statistical power to control for a range of potential confounders. The findings were also robust to a series of sensitivity analyses. An additional limitation that may affect the findings include diagnostic overshadowing, where a diagnosis of psychosis may decrease the likelihood of giving additional psychiatric diagnoses. Hence, the association between ASD and MTF may be an

underestimate. Another possible explanation for the observed association between ASD and increased risk of MTF could be that of misdiagnosis, where the association found between ASD and MTF could be explained by a subgroup within comorbid ASD that better fit a 'multidimensional impairment' phenotype, which I was unable to ascertain from the clinical record. Multidimensionally impaired children, first described by Kumra et al. 1998,[147] present with early transient autistic features, post-psychotic cognitive decline, and psychotic symptoms which are less likely to be amenable to antipsychotic treatment.[147,205] Another potential limitation is that individual reasons for each discontinuation of treatment were likely to be multi-factorial. By rating treatment failure to one of four potential categories at each point of discontinuation/failure, I may have underestimated the contribution of other underlying reasons. Nonetheless, inclusion of this additional information is likely to further support the study findings of the heterogeneity that underlies recurrent treatment discontinuation.

The study results are consistent with the evidence that shows psychotic illness experienced by children and adults with ASD may be different from non-ASD samples,[173,175] as I found diagnostic profiles in children with ASD comorbidity had lower rates of ICD-10 schizophrenia and higher rates of psychosis-NOS. Although there are risks of diagnostic misclassification between psychotic illness and ASD within the clinical sample, I believe the availability of detailed professional observations of children's behaviour within the free text records, has provided a greater clarity in the diagnostic validation of these complex symptoms, which are not always feasible using structured assessments.[146]

4.5.3 Conclusion

The findings provide evidence, at arguably the most sensitive point in psychosis development, that may help guide early detection of those children and adolescents at risk of not responding to first line antipsychotic medications. These findings may help delineate a subgroup of first-episode patients with EOP – i.e. those with co-morbid ASD - who have nearly double the risk for eventual development of MTF. This may explain why some children with premorbid difficulties and EOP are at increased risk for adverse social, educational, and occupational functioning.[169,206] Furthermore, given the size of the sample, the longitudinal nature of the analyses and the comprehensive review of psychotic symptoms within clinical text, I believe the findings provide further support for the atypical diagnostic distribution for psychotic illness in ASD previously described in both adult and child populations.[173,175] Further work could focus on identifying reliable predictors of response to non-dopaminergic treatments and adjunctive non-pharmacological interventions. This would enable stratified or individualised

treatment in specific patient subgroups, such as those children with psychosis and comorbid ASD. This could help direct finite resources to improve outcomes for those most in need, and reduce the current heterogeneity of therapeutic response. [207]

CHAPTER 5. NEGATIVE SYMPTOMS IN EARLY-ONSET PSYCHOSIS AND THEIR ASSOCIATION WITH ANTIPSYCHOTIC TREATMENT FAILURE

The contents of this chapter have contributed to the following:

Downs J, Dean H, Lechler S, Sears N, Patel R, Shetty H, Hotopf M, Ford T, Diaz-Caneja MD, Arango C, McCabe JH, Hayes RD, Pina-Camacho L. Negative symptoms in early-onset psychosis and their association with antipsychotic treatment failure (*Schizophrenia Bulletin*, *under revision*)

5.1 SUMMARY

Background: The prevalence of negative symptoms (NS) and their effect on prognosis for adolescents with a first episode of psychosis is unclear. In a sample of 638 adolescents with EOP (aged 10-17 years, 51% male), I examined the prevalence of NS at first presentation to mental health services, and whether NS predicted eventual development of antipsychotic multiple treatment failure (MTF) prior to the age of 18. (as defined in chapter 4: by initiation of a third trial of novel antipsychotic due to prior insufficient response, intolerable adverse-effects or non-adherence).

Methods: Data were extracted from the electronic health records held by child inpatient and community-based services in South London via CRIS. Natural language processing tools were used to measure the presence of Marder Factor NS and antipsychotic use. The association between presenting with ≥2 NS and the development of MTF over a 5-year period was modelled using Cox regression.

Results: Out of the 638 children, 37.5% showed ≥2 NS at first presentation, and 124 (19.3%) developed MTF prior to the age of 18. The presence of NS at first episode was significantly associated with MTF (adjusted hazard ratio 1.73, 95% CI 1.15–2.58; p=.008) after controlling for a number of potential confounders including psychosis diagnostic classification. Other factors associated with MTF included co-morbid autism spectrum disorder, older age at first presentation, and Black ethnicity.

Conclusions: In EOP, NS at first episode are prevalent and may help identify a subset of children at higher risk of responding poorly to antipsychotics.

5.2 INTRODUCTION

Early-onset psychosis (EOP), defined as onset before age 18 years, is a severely debilitating condition associated with long-term psycho-social impairment.[169] As a diagnostic term, EOP covers a broad range of psychiatric illness including schizophrenia spectrum, affective and other non-affective psychotic disorders.[123] Children with EOP often show significant levels of both positive and negative symptoms and disorganized behaviour. Relative to adult-onset psychosis, children and adolescents are more likely to have a background of longer durations of untreated psychosis, poor pre-morbid adjustment, and greater number of co-existing conditions, such as neurodevelopmental and substance abuse disorders.[164,168]

Compared to work examining the pathogenesis of adult and early-onset psychosis, studies which examine prognostic indicators in the years following treatment initiation are relative scarce.[169] From the research conducted, findings suggest that both a longer duration of untreated psychosis and poorer premorbid adjustment are associated with poorer recovery in EOP.[169] Despite previous evidence from adult-onset samples supporting the influence of negative symptoms (NS) on functional outcomes and recovery, the effect of NS on the prognosis of EOP remains relatively unexplored. NS symptoms include lack of motivation, problems with social interaction or diminished emotional range, and involve a loss or deficit in normal functioning.[208,209] They can be enduring and inherent to the core disease process (i.e. primary NS), or caused by other factors such as medication side-effects, positive symptoms, concurrent depression or limited social stimulation (i.e. secondary NS).[208,209]

At present it is difficult to assess the prognostic implications of NS at a young's person's first presentation with psychosis.[169] In adult-onset cases, NS are reportedly present at first-episode psychosis in about 30-50% of patients.[210,211] They are difficult to treat and are one of the main contributors to the functional disability observed in psychotic illness.[212–218] In EOP cases, NS are also reportedly stable over time, but little is known about the prevalence of these symptoms at first-episode.[219] Most studies so far have focused on early-onset schizophrenia,[176,220,221] which may not generalise to the heterogeneous population of young people that first present to child and adolescent mental health services. In addition, prior research findings have been limited by small sample sizes, convenience recruitment of more severe cases, or inclusion of those more amenable to taking part in a research study.[168,169]

In a large naturalistic sample of children and adolescents first presenting to services with EOP, I examined the prevalence of NS at initial contact with mental health services. To explore NS

as potential prognostic indicator, I examined whether NS at first episode predicted antipsychotic treatment failure. I measured treatment failure using a pragmatic measure, as defined by initiation of a third trial of novel antipsychotic (due to prior insufficient response, intolerable adverse-effects or non-adherence), which I called multiple treatment failure (MTF).[222] Previous work in adult-onset samples, suggests that NS characterize psychotic disorders with non-hyperdopaminergic pathophysiology,[199,223] which is supported by clinical evidence that NS in the first-episode are associated with poorer response to antidopaminergic effects of current antipsychotic treatment.[199,224] Therefore, I predicted that EOP patients with NS at presentation would be more likely to experience MTF. I also expected that this association would remain after taking account of potential confounders, including type of psychotic disorder, co-morbid depression, and additional markers of premorbid neurodevelopmental difficulties such as co-occurring autism spectrum disorders (ASD), hyperkinetic disorder and intellectual disability.

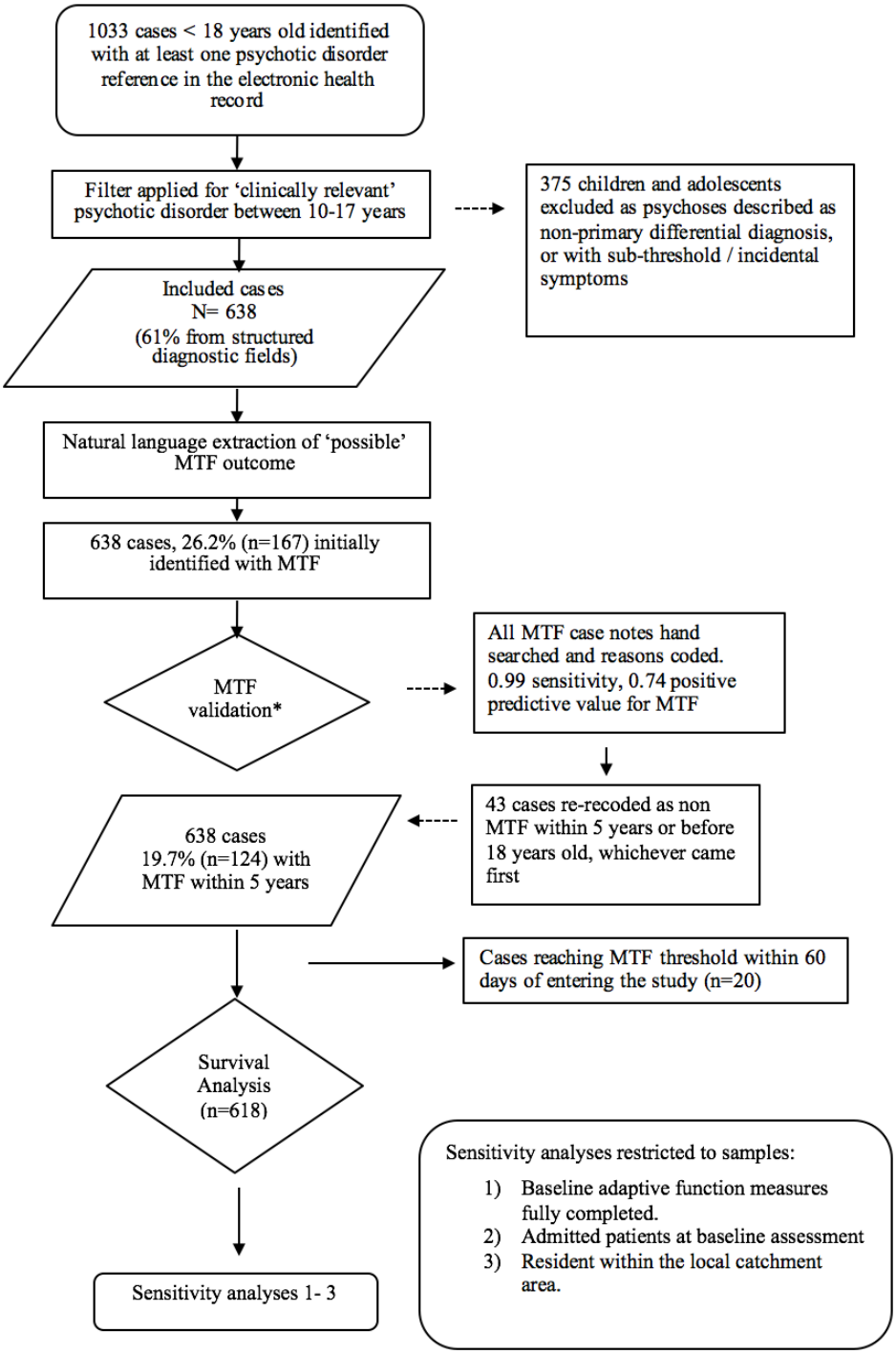
5.3 METHODS

5.3.1 Study design and study sample

A complete description of the study design and sample selection is provided in chapter 4. In brief, the sample consisted of a clinical cohort of all those individuals with a first episode of any psychotic disorder who were referred to SLaM CAMHS – including inpatient, outpatient and early intervention for psychosis services - between January 1st 2008 to December 31st 2014. Over this time, SLAM delivered all aspects of inpatient and community based child mental healthcare to approximately 280,000 children residing in four London boroughs, and specialist provision to children resident outside the boroughs where local area services (such as inpatient facilities) were unavailable. Most children experiencing a psychotic disorder within the SLaM catchment area of South London were likely to present to SLaM services and included in this study: the private sector has very limited involvement in child mental health within the area, and children with psychosis, relative to adults, usually come to the attention of services relatively early.[225]

As described in chapter 4, the included sample data were extracted using the CRIS application. Figure 5.1 shows the flowchart for inclusion in the study.

Figure 5.1 Flowchart for study inclusion and analysis



**Three independent raters, hand searched 100 cases each, including 167 cases where MTF was identified, and a random selection of non-MTF cases (44-45 per rater).*

Note: MTF: multiple treatment failure

Extraction of antipsychotic use data and definition of MTF

As described in chapter 4, [222] I used a previously validated GATE application to identify regular antipsychotic prescription trials from the structured medication fields and unstructured fields in the EHR.[132,160] Since no standard criteria for poor antipsychotic response or refractory disorder appeared suitable for EOP samples,[185,186] a proxy was created, based on the antipsychotic effectiveness literature,[187–189] which I termed MTF; defined as the initiation of a third trial of a novel antipsychotic due to insufficient response, intolerable adverse effects, non-adherence, or other miscellaneous reasons over a 5-year follow-up period from first presentation, or before the age of 18 years, whichever came first. Chapter 4 provided details around the validation of the MTF outcome and reasons for discontinuation.[222]

Extraction of NS data

A previously validated Natural Language Processing method [211] was used to find statements in the unstructured free-text fields of patients' EHR (i.e. progress notes, mental state assessments, discharge summaries, outpatient correspondence) which related to the presence of NS at baseline (i.e. within 60 days of accepted referral). The method was based on a NLP tool called TextHunter which has been described in detail elsewhere.[6] In brief, TextHunter is a custom-built NLP software tool which interfaces with CRIS. It facilitates each of the steps involved in developing a NLP application (previously described in the introductory chapter) from identifying appropriate ontologies and supporting manual annotation, to applying and testing sophisticated text based pattern recognition (including support vector machine learning approaches) derived from annotated training datasets.

A randomised sample of 100 cases was hand-searched by clinical raters, whilst blinded to MTF status. The PPV for NS subtypes ranged from 0.80 (poverty of speech) to 0.99 (mutism) and sensitivity ranged from 0.62 (poor motivation) to 0.97 (apathy). For the purposes of this study, Marder negative factor items[226,227] from the Positive and Negative Syndrome Scale (PANSS)[228] were used as a framework for characterising NS (see Table 5.1 for details). The extracted item 'social isolation' was considered descriptive of either passive apathetic social withdrawal (Marder N4) or active social avoidance (Marder G16). Having mutism, poverty of speech or both items recorded on the EHR was counted as a single NS, equivalent to lack of spontaneity / flow of conversation (Marder N6). The item psychomotor retardation (equivalent to Marder G7) was dropped as an NS due to its low PPV (0.55) and sensitivity (0.65). Furthermore, the hand search of the selected 100 cases revealed that this item had a low prevalence (~5% of the

sample) and always appeared acknowledged as an antipsychotic-related adverse effect (hence a secondary NS).

Table 5.1 Selection of negative symptoms from electronic health records and their equivalence to the Marder Negative Factor items within the PANSS

Items extracted from electronic health record	Marder Negative Factor items within the PANSS
Blunted affect	N1. Blunted affect
Emotional withdrawal	N2. Emotional withdrawal
Poor rapport	N3. Poor rapport
Social isolation	N4. Passive apathetic social withdrawal G16. Active social avoidance
Poverty of speech and/or Mutism	N6. Lack of spontaneity and conversation flow
Psychomotor retardation (dropped [a])	*G7. Motor retardation*

[a] Dropped from the study due to low PPV (0.55) and sensitivity (0.65) of the ´free text' extraction tool, and due to its being recorded mainly as secondary negative symptom.

Note: PANSS: Positive and Negative Syndrome Scale; PPV: positive predictive value

A composite ordinal variable, 'number of NS' (range 0 – 5) was created by summing the total count of the extracted NS. A score of at least two NS was applied a priori to determine the presence or absence of NS for analysis, and used to categorise individuals into having a positive NS profile (i.e ≥2 NS score) or non-NS profile. This approach was consistent with previous work that used the two-symptom cut-off to describe deficit syndromes in schizophrenia (i.e. primary, enduring NS).[211,216]

Extraction of other clinical and demographic data

A number of demographic variables and clinical data within 60 days of study entry (i.e. after accepted referral) were also extracted from the health record. Age at referral for first-episode psychosis, gender, ethnicity (according to categories defined by the UK Office for National Statistics), and index of neighbourhood deprivation for the main caregiver residence were extracted.[194] Data on illness severity and functioning around first presentation were extracted by means of the inpatient status and the Children's Global Assessment Scores (CGAS),[133] respectively. Data on ICD-10 co-morbid neuropsychiatric disorders which can be subsumed under the DSM-5 category of ASD (F84.0, F84.1, F84.5, F84.9), hyperkinetic disorder (F90.0, F90.1, F90.2, F90.8, F90.9), major depressive disorder (F32-33), and intellectual disability

(F70-79), were also extracted from free text and structured fields as previously described (chapters 2 and 4).[160,222]

5.3.2 Analyses

All analyses were conducted using STATA (Version 13). The prevalence of individuals meeting ≥2 threshold NS, and the total number of NS items was calculated. Logistic regression was used to examine whether NS profile was associated with demographic and baseline clinical characteristics.

To examine the prospective association between baseline demographic, clinical exposures and MTF outcome, I excluded children who had MTF within the 60-day baseline period (*n*=20). Kaplan–Meier curves were used to illustrate survival over time (probability of non-development of MTF), comparing those who were and were not presenting with ≥2 NS at baseline. After checking proportional hazards assumptions, I used a Cox regression to model the association between this baseline NS profile and MTF over a 5-year follow-up period from first presentation, or before the age of 18 years, whichever came first. The first model examined the crude effect of NS alone on MTF. Subsequent models were constructed adding potential socio-demographic and clinical confounders. As sampling bias towards more severe cases could affect the external validity of the findings, several sensitivity analyses were conducted to restrict the aforementioned models to (i) those children with complete adaptive function (CGAS) measures at first presentation (ii) inpatient children only; and (iii) those only resident within the local catchment area.

5.4 RESULTS

5.4.1 Demographic and clinical characteristics of the sample

Demographic and clinical characteristics of the 638 patients included (124 [19.3%] of whom developed MTF over time) and of the NS subgroup are presented in Table 5.2.

Table 5.2 Comparison between young people with early-onset psychosis at first presentation with and without ≥ two negative symptoms documented

Sample characteristics	Non-NS group (*n* = 399)	NS group (*n* = 239)	O.R (95% C.I)
MTF status, n (%)	59 (14.8)	65 (27.2)	2.15 (1.45-3.20)**
Female, n (%)	192 (48.1)	117 (48.9)	1.03 (0.75-1.42)
Age at referral (mean, SD)	15.4 (1.9)	15.9 (1.9)	1.17 (1.06-1.28)**
Age of reaching MTF (mean, SD)	16.5 (1.3)	16.0 (0.19)	0.79 (0.61-1.04)
Duration of follow-up (days), mean (SD)	721.4 (529.9)	590.5 (458.0)	0.995 (0.991-0.998)**
Ethnicity, n (%)			
White	204 (51.1)	93 (38.9)	Reference
Black	113 (28.3)	96 (40.2)	1.86 (1.29-2.67)
Asian	18 (4.5)	21 (8.8)	2.56 (1.30-5.03)
Mixed	47(11.8)	27(11.3)	1.26 (0.74-2.15)
Not Stated	17 (4.3)	2 (0.8)	0.25 (0.06-1.14)
Neighbourhood Characteristics, n (%) [a]			
1st (Least Deprived)	104 (27.1)	61 (25.9)	Reference
2nd	90 (23.4)	62 (26.4)	1.17 (0.75-1.42)
3rd	94 (24.5)	57 (24.3)	1.03 (0.66-1.63)
4th (Most Deprived)	96 (25.0)	55 (23.4)	0.98 (0.62-1.54)
First ICD-10 psychosis diagnosis, n (%)			
Other Psychoses	63 (15.8)	43 (17.9)	Reference
Bipolar Disorder	31 (7.8)	11 (4.7)	0.57 (0.24-1.15)
Drug-induced psychosis	29 (7.3)	10 (4.2)	0.51 (0.22-1.14)
Schizophrenia	222 (55.6)	143 (59.8)	0.94 (0.61-1.46)
Schizoaffective	11 (2.8)	6 (2.5)	0.80 (0.27-2.32)
Psychotic Depression	43 (10.8)	26 (10.9)	0.89 (0.47-1.65)
Co-morbid neuropsychiatric disorders, n (%)			
Autism Spectrum Disorder	75 (18.8)	39 (16.3)	0.84 (0.55-1.29)
Hyperkinetic Disorder	33 (8.27)	7 (2.9)	0.33 (0.15-0.77)**
Intellectual Disability	43 (10.8)	22 (9.2)	0.84 (0.49-1.44)
Major Depressive Disorder	108 (27.1)	66 (27.6)	1.03 (0.72-1.48)
Illness severity/ Functioning			
Admission at presentation, n (%)	90 (22.6)	170 (71.1)	8.5 (5.9-12.2)***
CGAS score (mean, SD)[b]	42.1 (15.3)	33.7 (15.4)	0.97 (0.95-0.98)***

*p< .05; **$p < .01$; [a] Variable dropped due to 0 values in cell. Note: O.R: Odds ratio; MTF: multiple treatment failure; NS: negative symptoms

5.4.2 Negative symptom prevalence

Table 5.3 shows the prevalence of each NS item using the manually-validated GATE extraction tool, in the total sample, and specifically for the MTF subgroup. Of note, 52.4% of the MTF subgroup presented with a positive NS profile. The most prevalent NS in the MTF subgroup was emotional withdrawal (43.6%). The prevalence of positive NS profile across diagnostic categories were as follows: schizophrenia- 39.2%, schizoaffective disorder- 35.3%, bipolar disorder- 26.1%, psychotic depression- 37.7%, drug-induced psychosis- 25.6% and other psychoses- 40.6%.

Table 5.3 Prevalence of negative symptoms at first presentation to services in early-onset psychosis subjects

NS items extracted from EHR	**Total sample (n= 638)**	**MTF (n=124)**
	n (%)	**n (%)**
Blunted affect	130 (20.3)	29 (23.4)
Emotional withdrawal	214 (33.5)	54 (43.6)
Poor rapport	62 (9.7)	18 (14.5)
Social isolation	51 (8.0)	14 (11.3)
Poverty of speech	32 (5.0)	8 (6.5)
Mutism	66 (10.3)	25 (20.2)
≥ 2 NS	239 (37.5)	65 (52.4)

Note: EHR: electronic health record; MTF: multiple treatment failure; NS: negative symptoms

5.4.3 Reasons for antipsychotic discontinuation

Details on the antipsychotic treatment pathways for the 124 children who developed MTF are shown in table 5.4. Cases identified as having the same reason for antipsychotic discontinuation at first and second antipsychotic trials were grouped into three MTF 'persistent reason' groups (persistent insufficient response, adverse events or non-adherence). A 'variability in reasons' subgroup (i.e. when reasons were different at each antipsychotic trial) was also created. The main patterns of discontinuation in the MTF group were the combination of insufficient response and adverse events (*n*=32, 35.2%), and persistent adverse events (*n*=19, 20.9%) over time. Children with NS profile showed higher rates of the 'insufficient response-and-adverse

effect' trajectory and lower rates of adherence-related trajectories relative to those with non-NS profile (table 5.4).

Table 5.4 Reasons for multiple treatment failure in young people with early-onset psychosis, with and without negative symptoms(NS) at first presentation

Reasons for MTF [a]	N (%) of individuals [b]	
	Non - NS (n = 41)	NS (n = 50)
Persistent insufficient response	6 (14.6)	7 (14.0)
Persistent adverse effects	9 (21.9)	10 (20.0)
Persistent non-adherence	2 (4.9)	3 (6.0)
Variability in reasons		
• Insufficient response and adverse effects	11 (26.9)	21 (42.0)
• Insufficient response and non-adherence	3 (7.3)	4 (8.0)
• Adverse effects and non-adherence	10 (24.4)	5 (10.0)

[a] Comparison in reasons for MTF between Non-NS and NS groups; χ^2 =4.39, df=5, p=0.49

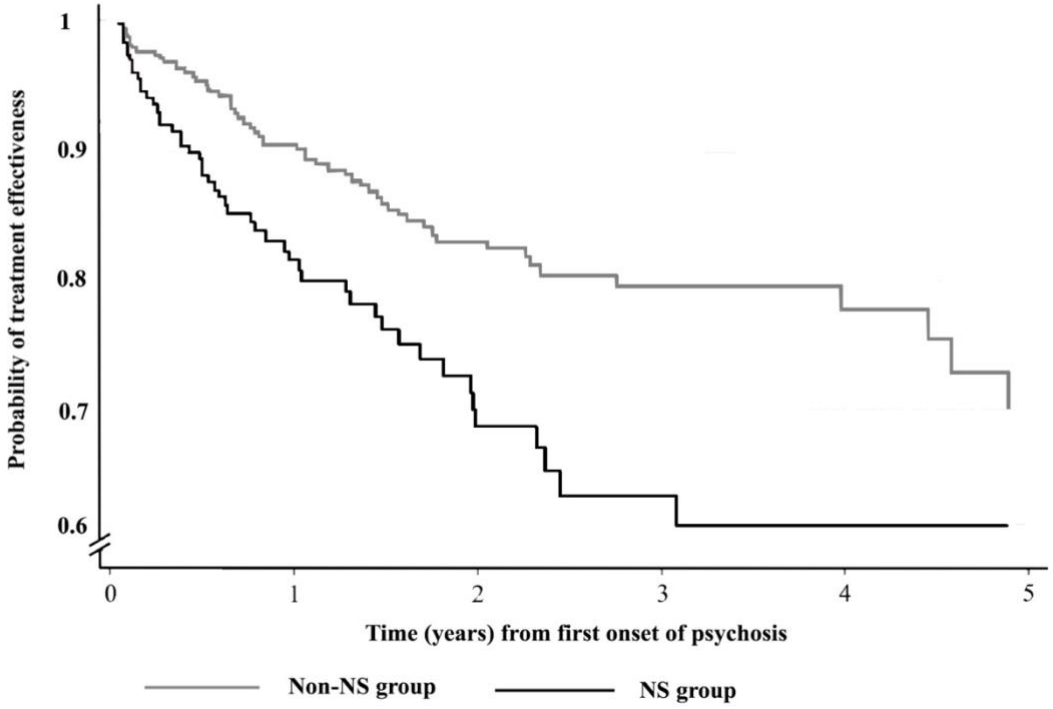
[b] In all cells, % refers to percentages (within columns) of individuals for whom information on main reason of discontinuation was available (n=91). Excluded due to no reason ' or 'other reason' ascertained were: Non-NS n= 18 (31%); NS group n=15 (23%)

Note: MTF: multiple treatment failure; NS: negative symptoms

5.4.4 Negative Symptoms and their associations with MTF

Kaplan-Meier curves displaying the survival status (probability of treatment effectiveness or non-MTF) over time of children with or without baseline NS profiles are presented as Figure 5.2. Those with non-NS profile at first presentation to services displayed significantly higher survival rate (p <.001). An adjusted Cox regression model (Table 5.5) revealed that NS profile was associated with increased risk of MTF over the follow-up period (adjusted hazard ratio [aH.R] 1.73, 95% CI 1.15–2.58; p= .008). Black ethnicity (aH.R 1.93, 95% CI: 1.17–3.03; p= .006), older age at first presentation (aH.R 1.29, 95% CI: 1.11–1.49; p= .001), and a comorbid diagnosis of ASD (aH.R 1.73, 95% CI: 1.05–2.83; p= .03) were also significantly associated with MTF.

Figure 5.2 Kaplan-Meier curves displaying the survival status (probability of treatment effectiveness or non-MTF) over time of children with or without negative symptom (NS) profiles at first presentation to services.



5.4.5 Sensitivity Analyses

A sensitivity analysis in all those with complete CGAS information (*n*=394), found NS profile was associated with increased risk of MTF (aH.R= 2.03; 95% CI= 1.18–3.48; p=.008). The analyses including only those individuals who were inpatients (*n*=260, 40.8%) at first presentation (within 60 days of accepted referral) or resident exclusively within the local catchment area (*n*=329), found little change in the direction and magnitude of the association between NS and MTF (aH.R= 1.68; 95% CI = 0.86–3.29; p= .13), and (aH.R= 1.67; 95% CI= 0.94–2.97; p=.08), respectively, although the reduced sample affected the power of the study to detect a significant association.

Table 5.5 Cox regression models for the association between negative symptom profile at first presentation and multiple treatment failure over time in early-onset psychosis (n=618)

Multiple Treatment Failure	Crude H.R. (95% CI)	Adjusted for socio-demographic factors H.R. (95% CI)	Fully-adjusted model H.R. (95% CI)
≥2 baseline NS	1.98 (1.35-2.91)**	1.66 (1.12-2.47)**	1.73 (1.15-2.58)**
Female (vs male) gender		1.08 (0.73-1.61)	1.19 (0.78-1.79)
Age at referral		1.25 (1.09-1.46)**	1.29 (1.11-1.49)**
Ethnicity			
White		Reference	Reference
Black		1.95 (1.23-3.00)**	1.89 (1.21-3.09)**
Asian		1.16 (0.48-2.77)	1.14 (0.47-2.76)
Mixed		1.51 (0.80-2.86)	1.12 (0.46 -2.72)
Not Stated		------[a]	------[a]
Neighbourhood Characteristics			
1st (Least Deprived)		Reference	Reference
2nd		0.60 (0.35-1.04)	0.69 (0.40-1.21)
3rd		0.55 (0.31-0.96)*	0.64 (0.35-1.09)
4th (Most Deprived)		0.55 (0.31-0.97)*	0.63 (0.35-1.10)
ICD-10 psychosis diagnosis			
Other psychoses			Reference
Bipolar disorder			1.65 (0.73-3.72)
Drug induced psychosis			0.95 (0.31-2.88)
Schizophrenia			1.17 (0.50-1.45)
Schizoaffective			2.57 (0.92-7.13)
Psychotic depression			1.32 (0.59-2.94)
Co-morbid major depressive disorder			0.64 (0.38-1.11)
Co-morbid autism spectrum disorder			1.73 (1.05-2.83)*
Other co-morbid neurodevelopmental disorder (hyperkinetic disorder / intellectual disability)			0.69 (0.38-1.25)

*p< .05; **$p < .01$; [a] Variable dropped due to 0 values in cell. Note: H.R.: hazard ratio; MTF: multiple treatment failure; NS: negative symptoms

5.5 DISCUSSION

This study shows that children and adolescents with psychosis commonly present with NS, with more than one third of the sample displaying NS at first presentation to services. The results also show that an NS profile at first stages is a prognostic marker for antipsychotic treatment failure in children with EOP: approximately 30% of the sample with NS at baseline went on to develop MTF, representing a two-fold increased risk from those without NS. The treatment pathway to MTF for young people with NS profiles appears to be driven by a combination of limited treatment response and emergence of intolerable adverse effects. Older age at first episode, Black ethnicity and a comorbid diagnosis of ASD are also significant predictors of MTF in this sample.

This is, to my knowledge, the largest naturalistic study of its kind to examine the prevalence of NS in EOP at first presentation to child mental health services. The study used an innovative text mining technique, adapted from an application in adult mental health records,[8] to extract negative symptom profiles. More than one third of the EOP population had two or more NS at baseline, rates that are consistent with those reported in both child and adult-onset psychosis literature (around 30-50%).[211,229]

This is also the first study to assess the association of NS and antipsychotic treatment failure in first-episode EOP patients. These results, combined with findings that NS can manifest in the psychosis prodrome,[230] suggests that NS profiles could represent a distinct phenotypic trajectory in young people with psychotic disorders. NS are possibly a marker for a distinct deviant neurodevelopmental trajectory which may be harder to treat with conventional antipsychotics and therefore result in a more impaired illness course. Although no previous work has examined treatment failure as an outcome in EOP, the findings are consistent with evidence that NS are associated with poor clinical outcomes in adult and child samples, many of those using validated gold-standard instruments to measure negative symptoms (e.g. the PANSS).[169,231] This work using text mining approaches for NS identification in large scale naturalistic samples of EOP using EHRs serves to complement the more traditional approaches using selective cohorts and intensive structured assessments, to inform prognostic indicators in clinical practice.

Several alternative psychopathological processes may be driving the study findings. Higher levels of primary NS may represent a clinical phenotype for greater levels of 'non-hyperdopaminergic' processes behind psychosis development.[199,223,232] Hence NS may help

identify a subgroup of patients with positive symptoms which do not respond well to antipsychotics, and at higher risk of developing MTF. Alternatively, NS may have an independent pathophysiology to positive psychotic symptoms, but may be moderating the association between positive symptom reduction and the protective factors required for a sustained remission.

The findings support the notion that NS are intrinsic to early-onset psychosis (across different psychosis diagnostic categories) and are already present during the first psychotic break. In regard to the prevalence across the different psychosis disorder classifications in this sample, NS were present in about one third of all EOP diagnostic subgroups, with slightly higher rates in those with non-affective psychosis. This suggests that in EOP, differences between psychosis diagnostic categories (especially between schizophrenia and affective psychoses) are quantitative rather than qualitative in nature, and all diagnoses are associated with presence of impairing symptoms (as reflected by similar rates of NS). Further studies using transdiagnostic approaches, as used in this study, are needed to advance the understanding of the physiopathology and predictive value of NS across disorders.

5.5.1 Strengths

The main strengths of this study include the use of a large sample of first-episode EOP, which provides a 'real world' sample of young people accessing inpatient and outpatient first episode psychosis CAMH services. Selecting an early-onset sample at first episode, reduces the potential bias incurred through unknown treatment exposures. The large sample size, and relative long duration of assessment provides sufficient power to estimate the association between NS and MTF even after adjustment for a number of potential clinical confounders, including psychotic disorder classification, neurodevelopmental and depressive disorder comorbidity. Using a clinical rater review of the whole electronic health record for sub-sets of patients allowed us to compute performance estimates of the different text extraction tools used in the study and select the most accurate ones, and enabled correction of misclassification errors.

5.5.2 Limitations

Results derived from this study should also be interpreted in the context of several limitations. Within the EOP sample, it was difficult to ascertain whether extracted NS were primary or secondary in nature, I assume that as NS were rated early (i.e. within 60 days of presentation to services and potentially prior or at the point of starting initial antipsychotic treatment), and

excluding the presence of psychomotor retardation from the total NS counting, the NS I detect, are mainly (but not only) primary in character.

In regard to the MTF definition, I was unable to obtain relevant antipsychotic data such as maximum daily antipsychotic dose, antipsychotic serum levels, or structured assessments of tolerability, which may have provided more objective assessments of treatment failure. Besides, by rating treatment failure to one of four potential categories at each point of discontinuation/treatment failure, I may have underestimated the contribution of other underlying reasons to treatment failure. As with all observational studies, the study findings may be limited by residual confounding, for example I was unable to adjust for the potential effects of substance misuse on MTF, and duration of untreated psychosis – both of which could be explanatory factors for older age being associated with MTF. Finally, there is a chance that not all children and adolescents experiencing a first-episode psychosis within the catchment area who access clinical services would have presented to SLaM CAMHs. Also given potential changes in residence away from SLaM services, it is possible that not all young peoples' psychiatric care was captured by the health record system over the course of follow-up. Given the mean duration of follow-up was lower in the NS group, I suspect that this may have led to an underestimation of the NS-MTF effect I report. Furthermore, the impact of potential loss to follow-up, and of non-actual first presentation to services, are likely to be limited, as I conducted a sensitivity analyses of children resident within the local catchment throughout the duration of their care, which showed little difference from whole sample findings.

5.5.3 Conclusion

In summary, this study demonstrated that there is a high prevalence of negative symptoms in early-onset psychosis around patients' first presentation to services and across psychosis diagnosis classifications, and supports the hypothesis that presence of these symptoms around the first stages of the illness identifies a subset of children who may be at higher risk of responding poorly to antipsychotics, both through refractory symptoms and high sensitivity to side-effects. Optimisation of current pharmacological and non-pharmacological strategies for these patients, and further research involving agents that better target negative symptoms are warranted.

.

CHAPTER 6. LINKING HEALTH AND EDUCATION DATA TO PLAN AND EVALUATE SERVICES FOR CHILDREN.

The contents of this chapter have contributed to the following:

Publication in a peer-reviewed journal

Downs J, Gilbert R, Hayes RD, Hotopf M, Ford T. Linking up data to plan and improve mental health services for children in England. *Archives of Diseases in Childhood* 2017;102: 599-602

6.1 SUMMARY

In this chapter, I provide an overview on the first area-based linkage in England which I conducted between mental health, hospital and school data, covering a total population of 1.25 million. I give an overview of this resource, give examples of how it is being used to improve public services for children, and discuss what is needed to implement this approach more widely across the UK.

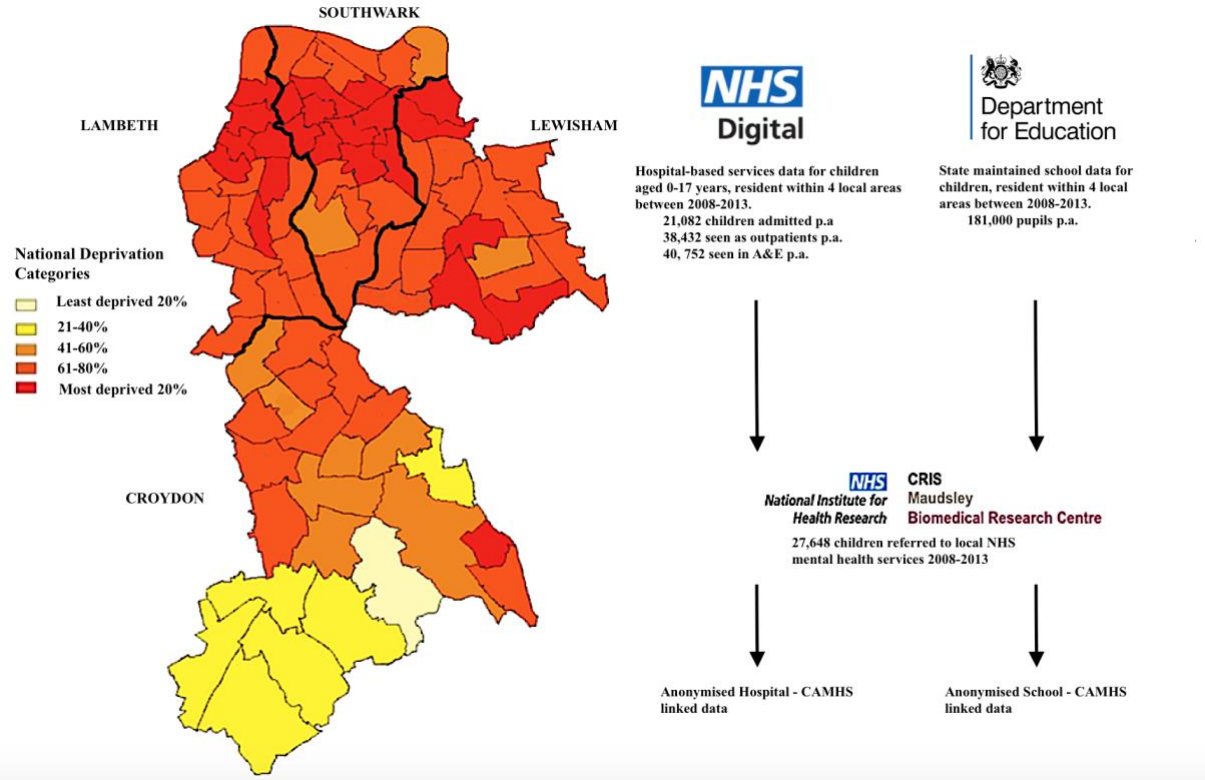
6.2 INTRODUCTION

Linkage of routinely collected data from public services has the potential to improve how local health, education, and social care are delivered to children. All mental health services, hospital-based child health services, schools and child protection services which serve the same local area, could be more efficient if the design, monitoring, targeting and integration of services were based on data. Health services need evidence from the populations that they serve to plan care and know whether they are meeting children's needs, duplicating effort, or allowing some children to fall through the net. In this chapter, I describe why I have joined up data from health, education and social services for children living in four local authorities in South London to create two datasets. One linking hospital to children's mental health services and the second linking mental health data to education data. I describe these resources, give examples of how they could be used to improve services, and discuss what is needed to implement this approach more widely across the UK.

6.2.1 What data are available?

Across England, all NHS health and state education services for children routinely generate administrative data, but few areas have managed to join these data systematically to evaluate how services could better serve their populations. Details of every NHS hospital inpatient admission, emergency department and outpatient contact are centrally collated by NHS Digital.[233] Demographic and socio-economic data on every child in state education are submitted by all state maintained schools to the Department of Education, along with information on school attendance, attainment, exclusion, child protection involvement, and special needs.[234] Centrally collected child mental health data has yet to become available, but nearly all local services collect these data within their electronic health record systems.[73] A big challenge is meeting the technical and governance requirements that safeguards sensitive child data, but also permits the linkage across public service data resources. This challenge has been addressed by the NIHR biomedical research centre at the Maudsley and there is the potential to extend our approach to other sites.

Figure 6.1 Linked data resources to provide an anonymised multiagency dataset covering child and adolescent mental health services, hospital attendances, education services and social service activity in South London.



As described in chapter 2, ten years ago, the Maudsley NIHR biomedical research centre set up the CRIS. CRIS has linked mental health data to education and hospital data. It took 3 years, from first application, to obtain permissions to do this from the Health Research Authority, NHS Digital and the Department for Education. I describe the legal, governance and technical challenges of this process in greater depth within chapter 7. In brief, the linkage process itself involved CRIS sending patient identifiers (names, and dates of birth, postcodes), without any mental health information to NHS Digital and to the Department for Education, where the identifiers were linked, and data from education and hospitals were de-identified and returned to CRIS. The CRIS secure environment now holds two linked datasets, education data linked to mental health data, and a second dataset containing mental health and hospital data. These datasets are kept separately, with all identifiers (names and NHS or pupil ID numbers) removed.

The CRIS system covers all NHS mental health services for four local authorities, which service a population of 1.25 million people. Patients using mental health services are made aware of how their data are used through notices in clinics, websites and regular public

engagement events. Although patients are not asked for consent to use their data for service evaluation or research, they are able to opt out. Only three individuals have asked to opt out of CRIS in six years. In 2014, the CRIS system was extended to four more mental health trusts (in 16 local authorities) and could be extended beyond mental health to other services.[96]

6.2.2 Using linked data from schools and mental health services

The population

The linked schools and mental health dataset captures data for approximately 160,000-190,000 children each year from 2007-13. To be included, children need to be aged between 4 and 16 years (see figure 6.1 for population numbers) and be resident in Southwark, Lambeth, Lewisham or Croydon. These areas are culturally and economically diverse, representing both outer and inner London regions. The catchment population has substantially higher proportions of families from black minority ethnic groups and/or born outside UK compared with rest of London and England. Highest and lowest socioeconomic groups are overly represented compared with England; with higher rates of unemployment, but also higher levels of education.[235] Linkage with the national pupil dataset means that information on education is still captured for those attending state school outside the local catchment area, and for those who move in or out of the area. Some of the population are not routinely captured. Children attending independent (meaning private) primary schools are not represented (~ 5% of the population aged under 12).[236] Children attending independent secondary schools are included when they sit any national examinations (e.g. GCSE or A level).

What can be measured?

Alongside socio-demographic characteristics, the national pupil dataset provides rich information on childhood development.[234] It tracks indicators of cognitive ability via routine teacher based assessment of language and numerical ability as children start school, and then via standardized academic assessments in mid and late childhood. It captures indicators of special educational needs such as physical problems, including deafness and visual impairment; emotional and behavioural problems, and autism spectrum disorder and learning disability. The dataset also captures episodes of children being excluded from school and indicators of absenteeism. Children's social care data has also been linked, which includes social service referrals and investigations, including details of children who are placed into out of home care.

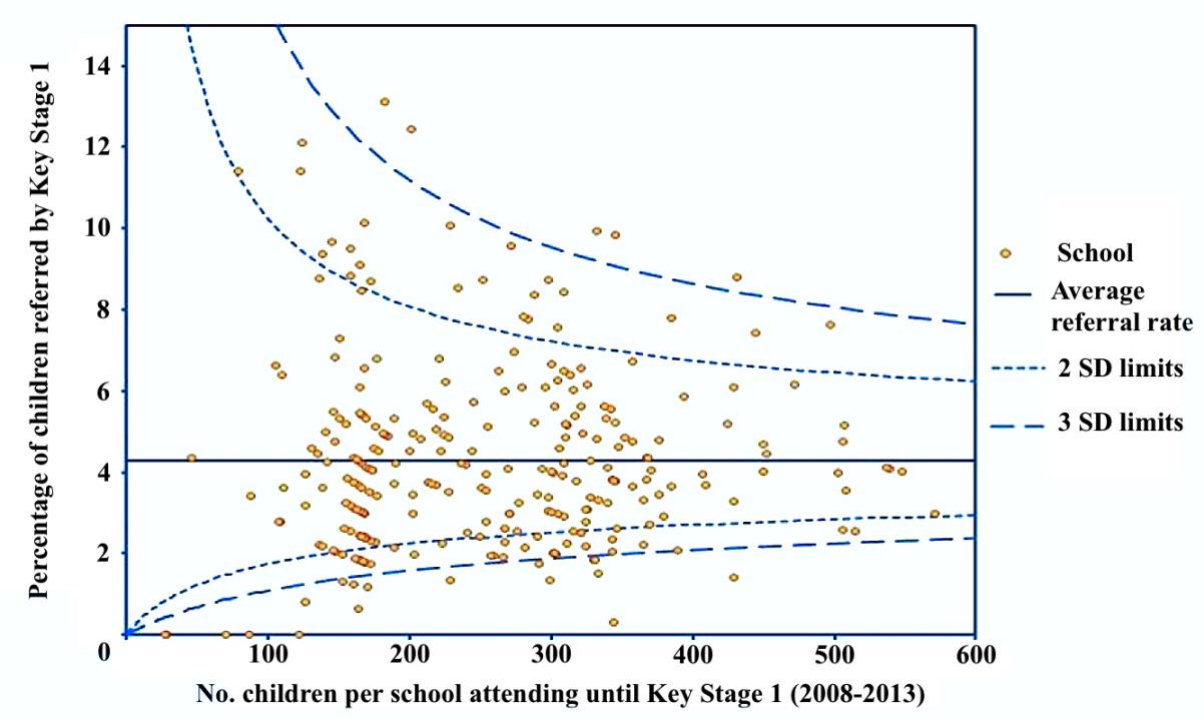
As described in chapters 2-5, the CRIS system enables researchers to access electronic mental health record data for approved studies.[127,129] Data available for research includes structured information (e.g. data entered by clinicians from drop down lists) such as past and present ICD-10 psychiatric diagnoses, appointments attended, and routine outcome measures (e.g. Strength and Difficulties Questionnaires)[126], and risk assessment details including risk of self-harm, self-injury, aggression to others.[160] Natural language processing software is used to enhance this data by extracting information predominately found in clinical progress notes and correspondence that might include more detail about family mental health problems, substance misuse, pharmacotherapy, and symptoms.[129]

How can school and mental health data be used to improve services?

The linked school and mental health data has many potential applications. It can provide detailed information on patient pathways and the extent of inequalities to services. This information can be used to flag gaps in existing healthcare provision and direct where new services are needed. National and local surveys have provided consistent evidence that timely access to services varies by social status and area of residence.[237] Data suggest that young people at high risk for mental health problems, looked after children and care leavers, those at risk of social exclusion or who have experienced abuse, or with long term physical health conditions, are the 'hardest to reach' and more likely to receive insufficient or fragmented care.[238] Because, local areas have considerable flexibility in how they commission child mental health services, there is a risk that 'hard to reach' groups are least likely to receive services. Mental health-school linked data can help understand which children receive support amongst socially vulnerable groups in each local area. Using data in this way to map service provision is particularly pertinent for integrated child health programmes which aim to tackle the potential inefficiencies and inequalities of current condition-specific pathways.[239] At present, local areas have very limited information on how mental health resources are accessed by vulnerable children, which include looked after children, those with a history of social services contact, prolonged absences from school[240] , permanent exclusions,[241] and with complex education needs.[242] Using these linked data, it is possible to gain a clearer picture of how well education and mental services overlap to address emotional and behavioural difficulties, and the shared awareness of special educational needs across both services.

Linkage of schools' data to mental health services also offers opportunities for targeting school based prevention strategies. For example, the funnel plot in figure 6.2 shows variation between mainstream schools in referrals to child and adolescent mental health services in the four local areas for children aged less than 8 years. Outliers on the funnel plots are of particular interest and warrant further exploration: very high rates could reflect high levels of population need and/or school-wide difficulties in managing emotional and behavioural problems, or conversely, very low referrals to mental health services may reflect excellent in-school support and provide a model of good practice. Using similar techniques, the data can be used to examine whether potentially more 'contagious' adolescent mental health problems like eating disorders, self-harm or suicidal behaviours cluster within schools. Findings can then be used to prioritise schools for preventive strategies. There is also a need for research to examine associations between educational achievement, self-harm presenting to mental health services, and the potential impact of school based interventions, as almost no research has been conducted on this topic in the UK.[243]

Figure 6.2 Plot showing referral rates to Child and Adolescent Mental Health Services for each school by Key Stage 1 (infant school)



Note: Funnel plot displaying referral rates as a function of the pupils enrolled between 2008 and 2013 within the school. The average referral rate is 4.3% (shown as a horizontal line). Control limits are also plotted above and below this mean.

6.2.3 Using linked hospital-mental health service data to inform services

There are a number of policy-relevant research areas that can benefit from using linked mental health and hospital administrative data. One example is the evaluation of policy initiatives to improve the quality of crisis care for young people.[244] There are approximately 200,000 episodes of self-harm that present to emergency services each year in the UK, with the highest rates amongst adolescents and young adults.[245,246] Between 25–50% of adolescents presenting to emergency care with self-harm do not attend any follow-up mental health support.[247–249] Emergency departments have an important influence on future engagement with treatment.[250] By adapting an approach developed in adult populations,[251] we can track temporal shifts in rates of emergency department attendances for self-harm or suicidal behaviour for the 40,752 children and adolescents seen each year from the four local authorities served by CRIS. We can assess whether practice changes in emergency departments result in reduced rates of attendance for self-harm in the long term.

Another example is the use of linked hospital-mental health data to follow up children hospitalised with long term conditions to investigate their use of mental health services and psychiatric co-morbidity. We can evaluate the types of patients who receive mental health care, when, and which factors are associated with treatment gaps and/or reliance on emergency care and unplanned admissions. These studies can provide information on whether systems of care need to change to ensure particular populations with chronic health problems, such as ethnic minorities or socially disadvantaged children, receive equitable access to mental health services.

6.2.4 CRIS: a sustainable resource for evaluating child health policy and service improvement

The CRIS system offers a sustainable resource for population-based analyses of linked patient level data to inform child mental health and acute hospital services and education services. Because CRIS uses data extracted from electronic record systems it provides a powerful platform for continuous evaluation of local child health policy initiatives.[252] The CRIS system provides an efficient, area-based resource for research, service planning and evaluation with patients followed up across the country. CRIS is being reproduced in other areas, potentially leading to a number of local areas having fine-grained information to better target local resources. However, there is still considerable work to be done. Health commissioners and other decision makers at local and national levels will need to develop sustainable means of implementing the knowledge which resources such as CRIS can deliver. This is essential if

we wish to complete the Learning Health System cycle (please see http://www.learninghealthcareproject.org), and use our informatics resources to drive healthcare improvement and innovation.[28] Alongside this, public engagement, understanding and support is vital. If we want to adopt these systems further families, child health advocates, academics, clinicians and policy makers will need to decide together how local linked resources are best safeguarded and used in commissioning services.

I hope in time that others will be encouraged to extend the CRIS model to link data for children across public services. By doing so I hope we will reduce the unmet need among vulnerable children and to move the discussions on from 'not knowing'[29,30] to accurate and responsive information on which to base public health strategies for children and young people.

CHAPTER 7. LINKING ADMINISTRATIVE DATA ON CHILDREN'S MENTAL HEALTH AND EDUCATION: GOVERNANCE, LEGAL AND TECHNICAL CHALLENGES

The contents of this chapter have contributed to the following:

Downs J, Ford T, Shetty H, Little R, Jewell A, Broadbent M, Deighton J, Mostafa T, Gilbert R, Hotopf M, Hayes R. (2016) Feasibility of Data Linkage: linking sources of child data to explore service utilisation and outcomes. UCL Child Policy Research Unit Report. Department of Health

Perera G, Broadbent M, Chang C-K, Callard F, Downs J, Dutta R, Fernandes A, Hayes R, Henderson, M, Jackson R, Jewell, A, Kadra-Scalzo G, Little R, Pritchard, M, Shetty H, Tulloch A, Stewart R. (2016) Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record derived data resource BMJ Open 6: 1-22 e008721

7.1 SUMMARY

Background: There are strong interconnections between public services which deliver health, education and social care for children. Improvement or withdrawal of any one of these services can help or harm delivery of the others. Because of their complementary nature, policy makers and service providers advocate that evaluations should involve linked routinely collected health and education data. Research using linked health, social and education data have been in place in Scotland and Wales for several years, but as of yet, no comparable linkage has been achieved in England. In this chapter, I present the governance, legal and technical challenges I encountered in achieving this link for four local authorities in South London, and review implications for future analyses by researchers and policy makers.

Methods: Approvals were sought from multiple government and ethical committees to link SLAM child and adolescent mental health service data to Department for Education (DfE) educational data held within the National Pupil Database. Under robust governance protocols delivered by the Maudsley BRC Clinical Records Interactive Search, and via an NHS trusted third party, I extracted the personal identifiers from the electronic health records of young people of a clinical cohort of all individuals aged between 4 and 18 years referred to NHS mental health care in England between 1st September 2007 and 31st August 2013. The DfE used combined fuzzy and deterministic approaches to match personal identifiers (names, date of birth, and post code) with NHS personal identifiers, and returned individually-matched educational performance records. The potential linkage biases using this process were evaluated by comparing socio-demographic and clinical characteristics between linked and unlinked SLaM cases. Methods to mitigate these biases and their impact on an important clinical factor-educational association (ICD-10 Axis One mental disorder and school attendance) were explored using linkage probability weighting and adjustment.

Results: Governance challenges included developing a research protocol for data linkage which met the legislative requirements for both section 251 of the NHS Act 2006 and The Education (Individual Pupil Information) (Prescribed Persons) (England) Regulations 2009(2). From a total 35,509 individuals referred to SLaM, 29,278 were matched to NPD school attendance records representing a linkage rate of 82.5%. There were significant

differences in sociodemographic, clinical and administrative characteristics between groups linked and not linked to school data. For example, children with a recorded ICD-10 mental disorder were more likely to have linked records compared those without ICD-10 disorder [adjusted Odds Ratio (aO.R) 1.11, 95% C.I 1.04-1.18]. Groups with a reduced likelihood of linkage included those first presenting to services in late adolescence (aO.R 0.67, 95% C.I 0.59-0.75) or having NHS address data recorded outside school census timeframes (aO.R 0.15, 95% C.I 0.14-0.17). No significant differences were found in linkage rates between children in the lowest and highest quartiles of deprivation (aO.R 1.03(0.92-1.15). ICD-10 mental disorder remained significantly associated with persistent school absence (aO.R 1.13, 95% C.I 1.07-1.22) after adjustments for linkage error.

Conclusions: It is feasible to link routinely collected education and health for most school aged children and adolescences at an individual level. However current linkage methods can introduce biases, with older groups who present to clinical services being less likely to be captured. Possible biases due to linkage error can effect risk factor-outcome associations and need to be addressed when analysing and interpreting results.

7.2 INTRODUCTION

As described in Chapter 1, large scale longitudinal cohort studies and clinical databases are essential tools for understanding the aetiology and outcomes of childhood mental and physical disorders, including rare or late adverse effects of treatments. However, maintaining the methodological quality of these studies is costly. For example, in the early 1990's the cost of setting up and sustaining the 15,000 families recruited to Avon Longitudinal Study of Parents and Children birth cohort study was around £1 million per year.[253] Furthermore longitudinal studies are rarely sufficiently resourced to sustain representation of their target population.[254] Sample attrition during follow up can introduce significant methodological biases and undermine generalisability.[44]

These challenges have led epidemiological researchers to consider alternatives to traditional data collection approaches, and use routinely collected information by public services. Taking the UK as an example, every school-age child now has a comprehensive digital record, which captures their contact with health, social and education services (see table 1.1, chapter 1). These include individual records of birth details,[233] school performance,[255] physical growth,[256] primary and secondary health care service use,[129,233,257,258] social and youth justice services contact,[255,259] employment and training.[260] Research initiatives in Wales and Scotland, have now created linked datasets derived from these data resources, and are using them to help direct local and national public health strategy.[261]

As described in the introductory chapter and chapter 6, the advantages of linking routinely collected child health and non-health data are potentially high. The process can extend the investigative range of longitudinal studies at relatively little cost, with no additional burden to study participants. However, these approaches also have limitations. Sample representation can still be lost in the record linkage process, especially when individual consent is required to link to additional data sources.[262] Attrition through non-response/consent to medical record linkage requests can lead to systematic differences between linked and non-linked samples. A number of studies linking health data, show ethnic minorities, lower socio-economic groups, and those with limited use or access to health service, are often underrepresented in studies using consent based approaches.[262] Crucially, data linkage studies, which have excluded samples based on non-response to linkage requests, risk losing the participants who matter the most to health researchers – the vulnerable groups where exposures and adverse outcomes are

most likely to aggregate.[263] Arguably, the opportunity costs are higher for children, as children have traditionally been underserved in research for a variety of reasons, including difficulties in recruiting adequate samples to investigate rare outcomes and their potential risk factors.[263]

There are other options available for child health researchers who wish to limit non-response/consent bias. A number of jurisdictions provide exemptions for the need to gain consent to link health records to other data resources. It has been suggested that these exemption routes, using alternative legal and governance frameworks may significantly limit bias normally incurred in consent based longitudinal studies.[44] These processes have their challenges. Certainly in England, the ethical and legal processes, as well as the technical security requirements, to gain exemption from individual consent for health data are stringent.[264] In England, large scale data-linkages using routinely collected health data via non-consent routes, have largely remained within the domain of NHS Digital. This is a national body, also known as Health and Social Care Information Centre (HSCIC), established in April 2013 by the Health and Social Care Act 2012, who are responsible for centrally collecting, analysing and disseminating health and social care data submitted by NHS Trusts.

As yet, the potential gains from these 'big data' systems to drive local population-based analyses for child public mental health and educational services improvement remains unrealised. This chapter shows that it is possible for an individual NHS trust (South London and Maudsley NHS Foundation Trust) to create linkage environments that conform to NHS safeguards within England, and develop sustainable research systems that link and anonymise individual children's records from healthcare, social and educational systems. Expanding on the overview provided in chapter 6, I show how a linked resource between CRIS and the NPD [255] was created to provide whole-region population longitudinal dataset of childhood mental health disorders and educational outcomes. I describe the data preparation process and methodological approach taken to overcome the lack of a shared identifier number between health and education, in order to best link partial identifiers held on both datasets.

In the first part of the results section, I describe the challenges of gaining approval for a research protocol which needs to meet the legislative requirements for both section 251 of the NHS Act 2006, via recommendation from NHS Health Research Authority Confidentiality Advisory Groups, and The Education (Individual Pupil Information) (Prescribed Persons) (England) Regulations 2009(2).

In the second part of the results section, I provide an evaluation of the socio-demographic and diagnostic factors associated with the risk of non-matched health and educational records in the sample. As even when consent is not required for linkage, data matching processes can add non-linkage bias - a type of sampling bias that occurs when subjects are excluded because their linking variables do not adequately match between data sets - generating differences between those who are linked and non-linked. Errors in linkage may occur where there is no unique identifier across different data sets.[265] Of the few studies conducted which have examined linkage error between large scale datasets without shared identifier codes, their findings suggested that such biases derived from these linkage processes can be substantial,[266] and crucially, incomplete data linkages can result in systematic bias in reported clinical outcomes.[267] Fortunately there are a number of statistical approaches commonly used for reducing the potential selection bias incurred through non-linkage, such as inverse probability weighting and match probability adjustment.[268,269]

Linkage error may be particularly prominent in the DfE and SLaM health records as both use different identifier codes (a DfE pupil ID, and NHS number respectively) so linkage is based on matching on personal information such as name, sex, date of birth and postcode. The SLaM-DfE linked database was built for the purpose of conducting a number of observational studies which test hypothesised risk factor and outcome associations between mental disorders and school performance. However, the potential findings may be severely limited if linkage biases are not accounted for. An aim of the work described in this chapter was to use the linked data resource, and conduct several exploratory analyses to examine how potential linkage biases may impact potential associations between child health factors and school outcomes, in this case, school absence.

School absence was chosen as the outcome to assess linkage error because is it challenging to assess the impact of the error for a particular outcome, when there is not an expected one-to-one relationship between one variable and another. For example, when linking patient records to a death registry to determine a patient's survival status, it is difficult to know which matches have been missed – the death registry will only contain patients who have died, and so a non-match could be due to patient being alive or being a missed match.[262] Applying this to school data, there is a need to select a clinically relevant school performance outcome which should be available for all pupils. School attendance should be recorded for all pupils, and is clinically relevant, hence it is useful as outcome for evaluating the impact of linkage error.

Using the linked data, a cross-sectional study was conducted to examine the association between child mental health disorder and persistent school absence, with adjustment for potential linkage bias.[268,269] Changes in the main effect estimates of mental disorder on school absence were examined before and after adjustment for non-linkage bias, to determine the potential influence of linkage error on these associations.

7.3 METHODS

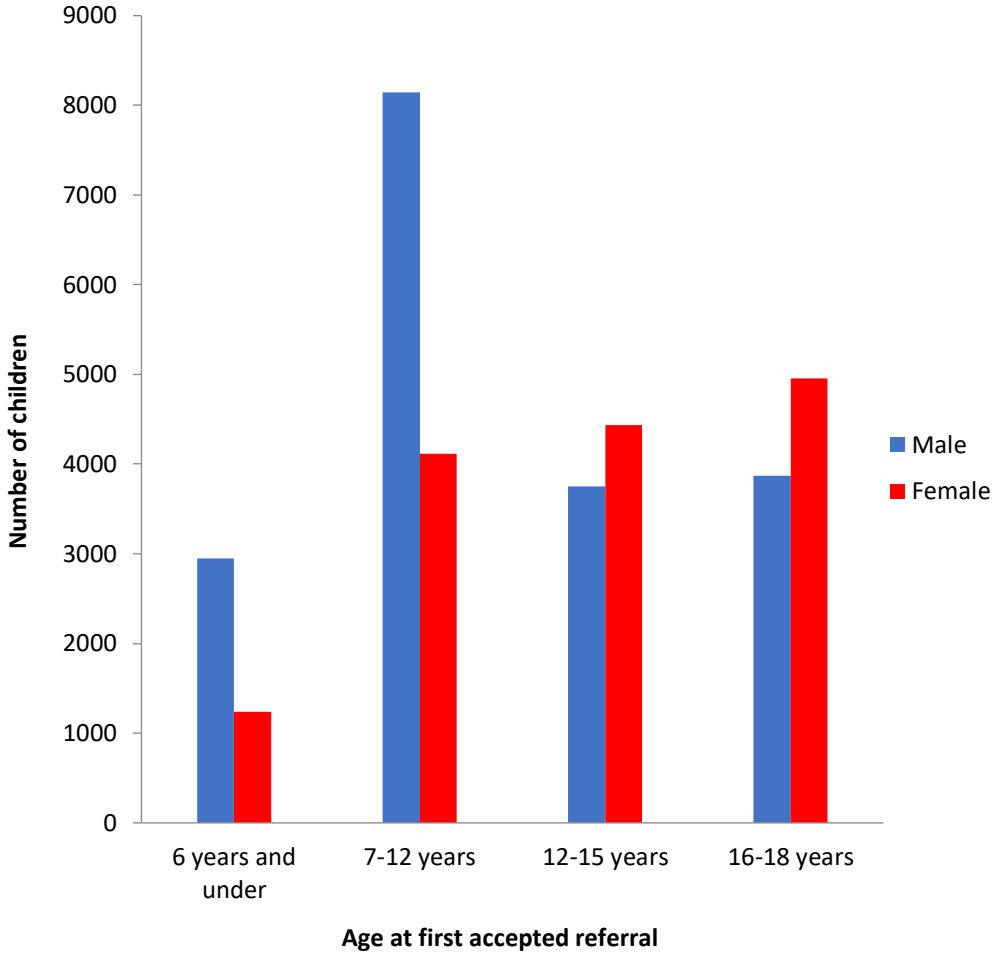
7.3.1 The data resources

NHS Child and Adolescent Mental Health Service Data

As described in previous chapters (2-6) SLaM provides comprehensive CAMHS to a geographic catchment of over 260,000 children in four south London boroughs— Croydon, Lambeth, Lewisham and Southwark— as well as some specialist services which also accept referrals from outside the four-borough catchment area. Figure 7.1 illustrates the number of children, by age and gender first accepted into SLAM CAMHS over a 5-year period. SLAM has dedicated multidisciplinary services for children, which assess and treat school age children with suspected or previously confirmed psychiatric disorders. The majority of children referred to CAMHS will be assessed under diagnostic criteria using the ICD-10 multi-axial classification system (See table 7.1).

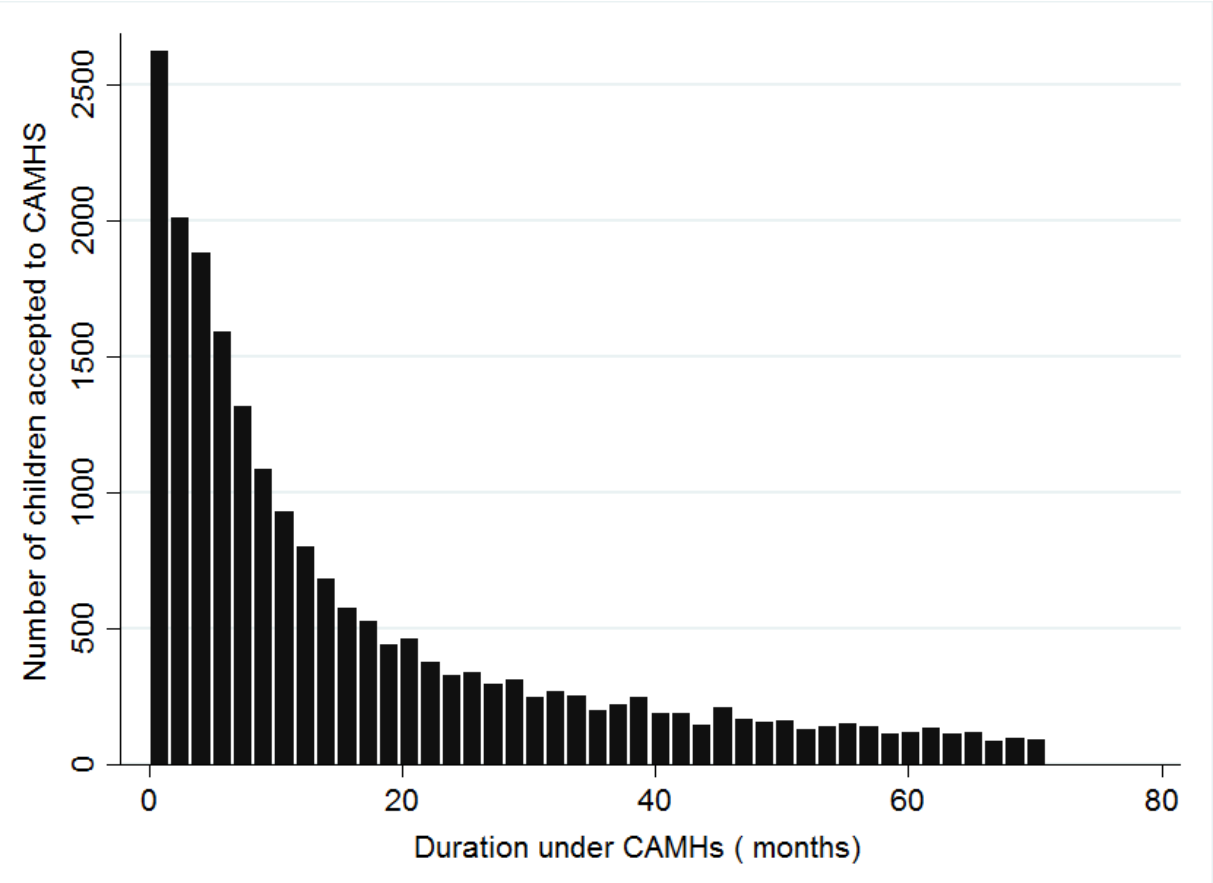
Children seen by CAMHS in the SLaM catchment are referred from primary care, child health, and educational and social care services, and typically undergo a multidisciplinary assessment by CAMHS clinicians. As shown in Figure 7.2, the majority of young people accepted into CAMHS receive short discreet periods of care, however some will receive multiple episodes of CAMHS support throughout their childhood. Clinical records have been fully electronic (i.e., paperless) across SLaM services since 2007.

Figure 7.1 Number of accepted first referrals for all children (aged 4 -16) seen by SLaM CAMHS services (Sept 2007 – August 2013)



CRIS extracts information from the records generated by CAMH services. CRIS has been described in detail in chapter 2-6, and elsewhere.[127,129,160,270,271] Because of the inclusion of both structured and unstructured (open-text) data in anonymised form, CRIS is unique in the depth of information that can be utilised in comparison to other case registries across the world.[65]

Figure 7.2 Duration between first and last contact with mental health professionals for children (aged 4 -16) accepted to SLAM CAMHS between Sept 2007 – August 2013.



Department for Education National Pupil Database

As described in chapter 6, the NPD is a pupil level longitudinal database which matches pupil and school characteristic data to pupil level attainment.[255] The key datasets within the NPD are the pupil census and pupil attainment datasets, which holds data for all assessments that pupils complete during primary and secondary school state education. The census is a snapshot of pupils attending maintained schools in England, which is submitted on a specific day by a school for all pupils in that school. It has been comprehensively collected across English schools since January 2002. It contains characteristics such as names, age, ethnicity, addresses, school details, special educational needs status and free school meals status. It also collects additional information including primary language spoken at home, termly school attendance and exclusions, social care involvement. These data have been submitted to the DfE via the schools' Management Information System. Pupils held within the NPD are typically aged between 3-19 years, but some from special schools may be up to age 24.

The pupil attainment dataset holds pupil-level longitudinal Key Stage attainment records. The first attainment assessments completed are for Early Years Foundation Stage profile when pupils are in reception aged between 4 and 5. They then complete Key Stage 1, Key Stage 2 and Key Stage 3 assessments when aged 7, 11 and 14 respectively. Key Stage 4 (GCSE) and Key Stage 5 (A-Level) assessments are typically taken when aged 16, 17 and 18. Key Stage 2 data has been collected since 1997, Key Stage 4 and 5 since 2002, and Key Stage 1 since 1998. As with the census, pupil attainment data are tracked across schools and charted throughout their school careers. The NPD also provides characteristics of the school. There is scope for linking the data from other related datasets such as national higher education databases (HESA), or teacher surveys.[255] The data collected has evolved over time, especially in relation to education attainment. This reflects the dynamic nature of DfE policy in relation to the timing and measurement of children's educational progress throughout their school career.

Table 7.1 Diagnostic breakdown of all children (aged 4 -17) referred to SLaM CAMHS services between Sept 2007 and August 2013.

ICD-10 Psychiatric Diagnostic Classification		Local Catchment Area*		National Catchment Area*	
		Male (n=15204)	Female (n=11469)	Male (n=4522)	Female (n=4314)
		n (%)	n (%)	n (%)	n (%)
	Any ICD-10 Diagnosis	9315 (61.3)	6587 (57.4)	2592 (57.3)	2545 (59)
Axis One	Pervasive Developmental Disorders (F84)	2116 (13.9)	519 (4.5)	749 (16.5)	248 (5.9)
	Hyperkinetic Disorders (F90)	2345 (15.4)	435 (3.8)	801 (17.7)	210 (4.9)
	Conduct Disorders (F91)	2160 (14.2)	983 (8.6)	392 (8.7)	169 (3.9)
	Disorders due to psychoactive substance use (F10–F19)	253 (1.7)	180 (1.6)	112 (2.5)	53 (1.2)
	Psychotic Disorders (F20-F29, F30-F31, F32.3)	437 (2.9)	438 (3.8)	239 (5.3)	239 (5.5)
	Depression and other (affective) disorders (F32–F39)	733 (4.8)	1497 (13.1)	197 (4.4)	511 (11.8)
	Emotional and stress related disorders (F40-F48, F93, F94, F98)	2442 (16.1)	2930 (25.5)	522 (11.5)	879 (26.4)
	Post-Traumatic Stress Disorder (F43)	269 (1.8)	330 (2.9)	64 (1.4)	105 (2.7)
	Obsessive Compulsive Disorder (F42)	201 (1.3)	220 (1.9)	269(5.9)	164 (3.9)
	No recorded Axis One Diagnosis	5889 (38.7)	4882 (42.6)	1929 (42.7)	1770 (41.0)
Axis Two	Disorders of Scholastic Development (F80-F89)	1048 (6.9)	337 (2.9)	195 (4.3)	89 (3.1)
Axis Three	Intellectual Disorders (F70-F79)	870 (5.7)	357 (3.1)	443 (9.7)	195 (3.8)

*Note: The sample are split by residence, either within 4 London Boroughs served by local SLaM services (Local Catchment area), or from rest of England served by SLaM National and Specialist services (National Catchment Area).

7.3.2 The technical resources

To link CRIS data with other external clinical and non-clinical sources, SLaM has developed a research governance model for linking data which satisfies NHS requirements as described in Department of Health Information Governance Review, or Caldicott 2, report.[272] In accordance with these guidelines, SLaM set-up the Confidential Data Linkage Service (CDLS),[129] which acts as Trusted Third Party or Safe Haven to ensure that confidential patient information can be linked in a way that guarantees the legal and ethical rights of patients. For the purpose of this linkage, a similar provision was available in DfE Data Services Provision, which had a linkage service, governed under HMG Security Policy Framework v10 2013 (SPF),[273] with experience of regularly undertaking external linkages with large scale research cohorts including the Millennium Cohort Study and ALSPAC.

7.3.3 Linkage

Preparing the CRIS CAMHS identifiers for matching.

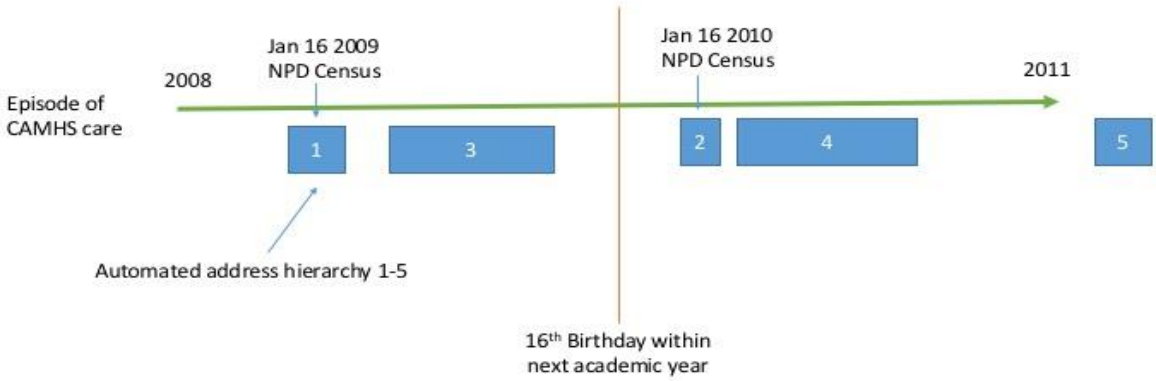
We selected a cohort of young people aged between 4 and 18 years, who were referred to SLaM mental health care between 1st September 2007 and 31st December 2013. As described previously, in the UK, unique identifiers, such as national health identifiers, are not shared between health and education databases, so records require matching on personal identifiers common to both data resources (i.e. names, dates of birth, and residence post code).

Personal identifiers were standardised using the following definitions:

1. **Dob**: format (dd-mm-yyyy)
2. **forename_1**: The first word present in the forename field registered for the individual record. (i.e. all text left of the first white space character in the free text field)
3. **forename_2**: The second word present, if >1 forename present (i.e. second of 2+ names separated by one space or punctuation except "-") (i.e. right of white space)
4. **surname_1**: The first word present in the surname field registered for the individual record. (i.e. all text left of the first white space character)
5. **surname_2**: The second word present, if >1 Surname present (i.e. second word of 2+ names if separated by one space or punctuation except"-")
6. **surname_3**: The whole string in the surname field

Within the longitudinal health record, there were often several different addresses held for each individual. Similarly, there were multiple addresses held for most pupils in the education database. Pupil address data are routinely updated on the 16th January every year. So, I developed a hierarchical system to extract the postcode from health record most likely to match with education database. Figure 7.3 shows how this postcode hierarchy might be applied to one individual child, where the blue blocks represent episodes of care provided by CAMHS, and the green time line represents the period of time in school. Taking these considerations into account I produced a hierarchy of postcodes with 1 to 5 levels for each individual seen in CAMHS using logic rules (see figure 7.3 legend).

Figure 7.3 Creating a hierarchy of matching postcodes* to improve the link between CRIS CAMHS Data to DfE National Pupil Database



Legend

1 Address most likely to coincide with the school census that we have recorded before the child is 16.

2 Address most likely to coincide with the census that we have recorded before the child is 18.

3 Address held for the longest duration by the child that we have recorded before the child is 16.

4 Address held for the longest duration by the child that we have recorded before the child is 18.

5 Any available postcode where 1-4 not available

*Note: the numbers within the blue block and the corresponding legend in figure 7.3 represent the respective postcode hierarchy category

A SQL based query was used to extract the identifier data according to these rules. This produced a sample of 36,760 individuals with distinct individual records. Post extraction I then ran data cleaning and logic checks which included removal of all those with numbers in name string fields (4 case removed), all those with only one letter in their first or surname (1 case removed), all those with incomplete / atypical English postcodes (214 records hand searched,

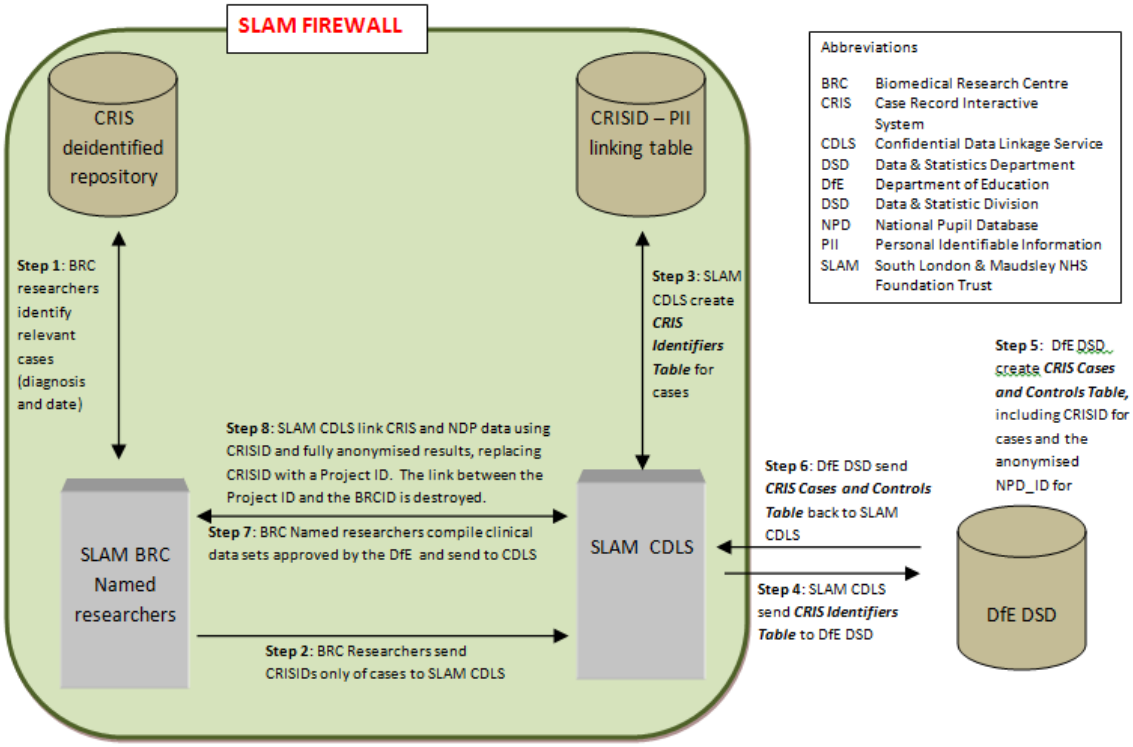
77 valid English postcodes were cleaned and retained). We excluded all children whose first referral date was less than 4 years (1095 days) after their Date of birth, unless they had confirmed follow up contact details recorded within the window (i.e. 2007-2013) at least one year later than the earliest referral date. This was because clinicians can erroneously record the date of referral or time seen at initial appointment in the date of birth field. This mainly occurs in individuals with only single episodes of contact with services. To fit in with the academic calendar and UK school age, children were then selected if they had their 4th birthday prior to the 1st September 2012. This provided a complete sample of 35,509 ready for matching with the NPD.

All the data prepared for matching had personal identifier fields populated with the exception of the secondary surnames and forenames (i.e. there were no missing values). Dates of Birth ranged from 06/01/1989 – 31/08/2008 which meant that all of these pupils could potentially be found in either current or historic NPD census data. Personal identifiers were standardised to maintain a consistent format with NPD identifiers: SLaM identifiers were prepared to fit with DfE first name, surname and date of birth formats, which included standardising string length, capitalizations, use of spaces, and hyphens.

Only identifiers (names, postcode and date of birth), accompanied by their unique CRIS ID pseudonym, were then sent via secure file transfer to the DfE Data and Statistics Department.

As represented in figure 7.4, the DfE matched these against NPD personal identifiers (approximately 15 million records), generating a pupil-specific, non-identifiable NPD ID variable across the whole data set, and adding the CRIS ID to this table for cases only, stripping the resultant table of all identifiers other than the anonymised NPD ID and the pseudonymised CRIS ID, and transferring the data set back to SLaM CDLS using a secure file transfer.

Figure 7.4 Data flow process linking CRIS CAMHS Data to the National Pupil Database



The supplied data items by the CDLS were matched to the NPD data by DfE informaticians in the stages described below. Initial matching or stage 1 was based on exact matches for the supplied data items. For those cases who did not match at stage 1, stage 2, 'fuzzy' matching processes were conducted, and so on, down to stage 4.

- Stage 1: Full match on names (all supplied values including alias), dates of birth and postcode (all supplied) were conducted against all years/terms of the School census data, Pupil Referral Data, Alternative Provision Data, Early Years Census data. School census data contained preferred and former surnames, which were also searched. Forenames were checked against forename/middle name combinations.

- Stage 2: Full match on Date of Birth, Postcode and Fuzzy matching on names. To ensure confidence in these matches, results were checked manually. Fuzzy matching was conducted on first two characters of names.

- Stage 3: Full match on names and dates of birth, postcode inward code (the first 2-4 characters) plus first character of the outward code (the latter characters after the space). To ensure confidence in these matches, results were checked manually.

- Stage 4: Full match on names and Postcode with manual check of dates of birth, looking for 'near' dates of birth – where the record may be possibly one year out, one month out, one day out, and transposed month/day.

7.3.4 Analysis of linkage bias

Overall linkage rate was calculated as the percentage of CAMHS individuals linked to any NPD school record on any of the stages 1-4. Potential sources of linkage biases were estimated by comparing linked and unlinked data. For the CAMHS sample described in table 7.1, I categorised an individual match to NPD school absence data (a subset of the NPD school record) as a binary outcome: match =1, non-match=0. I used the ICD-10 multi-axial classification system, to categorise the presence of any recorded mental health diagnosis (i.e. diagnoses status prior to 18th birthday) available between 2007 and 2013.

Using logistic regression, I explored the associations between a number of risk variables including demographic (e.g. gender, ethnicity, neighbourhood deprivation), clinical (age at first presentation to CAMHS, diagnosis of any ICD-10 disorder) and administrative factors (e.g. postcode hierarchy) with linkage to the school attendance database as the binary outcome. We used this logistic regression to generate a probability of matching estimate as a function of the risk variables.

7.3.5 Analysis of linkage error using school attendance outcomes

For each matched CAMHS-NPD pupil, I created a binary outcome marker of poor attendance for the latest academic year they attended school available between 2007/08 and 2012/13. I categorised pupils as poor attenders if they had recorded less than 80% school attendance for the total number of possible school sessions available since their enrolment for that academic year (one session is equal to half a school day).

Using the probability of matching estimate from the linkage bias analysis, I created a weight that was inversely proportional to the probability of being linked to national pupil database school attendance data, which I assigned to each individual with linked CAMHS-school absence data. This followed standard methodology for managing non-response bias in conventional cohort and survey designs.[268] I then ran a multivariable logistic regression using

the same predictor variables to examine their association with persistent school absence, initially without weights, and then with inverse probability weights. To examine another approach to adjust for potential selection bias from non-linkage,[269] I examined whether the main effects of interest also persisted after the probability of matching estimate was entered as a covariate in the multivariable logistic regression model.

7.4 RESULTS 1

7.4.1 Outcomes from linking the health and educational data resource: achieving the ethical, governance and legal approvals

The proposal to link the NPD and CRIS CAMHS data, underwent a robust and lengthy ethical, legal, governance and technical review, conducted by a number of local and national committees within NHS and DfE. Figure 7.5 provides the timeline and milestones achieved to reach the completion of the linked DfE-SLAM CAMHS dataset.

Gaining the permissions to link the NPD and CRIS CAMHS data was complex. There was no precedent in England for such a linkage between routinely collected mental health and school data, and there had been no successful completion of linked NHS and non-NHS non health data without individual consent.[274] I approached the Department for Education directly who held nationally collected education data via termly school submissions to the National Pupil Database.[255] I planned a linkage with national data, as opposed to regional data sources held by the local education authorities, to prevent clinical sample attrition. I expected a considerable proportion of children receiving SLaM treatment would reside outside the SLaM Catchment area or potentially move outside the catchment after treatment. In addition, the Department for Education had relatively transparent systems, and a dedicated office, for managing requests for educational data extracts, through their National Pupil Database Team. Once Research Governance approval was granted by the SLaM Caldicott Guardian Committee and the DfE's Data Management Advisory Panel, I submitted an application to the Health Research Authority Confidentiality Advisory Group (HRA CAG).[264] The HRA CAG have the authority to provide recommendations on behalf of the Secretary of State for Health, to permit the linkage of NHS data, without individual patient consent, for the purposes of research, if it meets the criteria within section 251 of the NHS Act 2006.

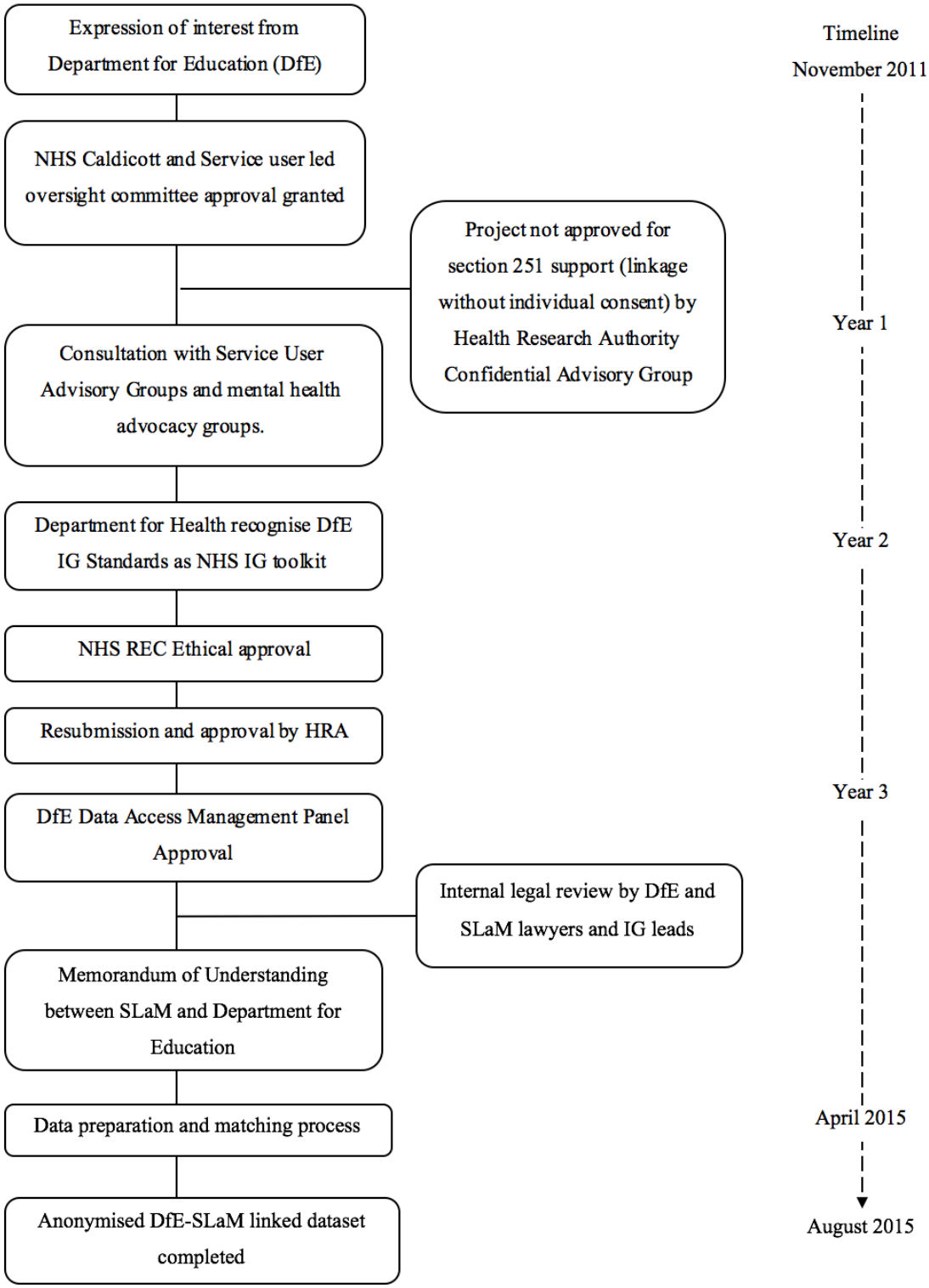
The HRA CAG rejected my first application, as the research activity proposed did not demonstrate sufficient medical purpose and public benefit to meet the s251 requirements. It was highlighted by the HRA CAG that support under current regulations could only be provided where potential public benefit were sufficiently defined (see HRA CAG guidelines for a general discussion on what research constitutes being for 'public benefit'[275]). In particular, it was noted that in order to satisfy one of the conditions in schedule 3 of the Data Protection Act [276] (required to process sensitive personal data including data relating to an individual's physical or mental health) a medical purpose would also need to be specified. A second issue, was the lack of consideration of a practicable alternative to the use of confidential patient information without consent.

The HRA CAG also queried whether I had considered if the Health and Social Care Information Centre (HSCIC, now NHS Digital[277]) could carry out the linkages on the applicant's behalf using their Trusted Data Linkage Service. The CAG advised that this route would negate the requirement for SLAM to disclose confidential patient information to the DfE, and minimise the disclosure of patient information. A final major issue related to the governance arrangements in place around the processing of patient data by the DfE. I hadn't provided sufficient information around retention periods, access arrangements and the extent of identifiable data requested.

7.4.2 Defining 'medical purpose' and public benefit when seeking s251 support

To prepare for resubmission, I examined the issues identified by the HRA CAG. My initial application took a broad interpretation of 'medical purpose.' Given my experiences as CAMHS clinician, and the time CAMHS devoted to improving children's function in school, I had presumed that educational outcomes for children with psychiatric diagnosis were salient to 'a medical purpose.' As a result, I had underestimated the need to demonstrate to the CAG that educational performance (attainment, attendances and exclusions) were viewed by researchers, and NHS clinicians working with children with mental disorders, as key medical outcomes. Also, I had not made a clear enough case for using the linked educational data to examine the aetiological factors for child onset psychiatric disorders. These issues were addressed in the revised scientific proposal, largely by describing research that would examine the bi-directional associations between educational performance and mental health disorders.

Figure 7.5 A timeline of the ethical, legal and technical milestones for reaching a data linkage between DfE and SLaM



In terms of gathering evidence for support of the public benefit to use patient identifiable data via CRIS to link to the national pupil database without patient or caregiver consent, I consulted several clinical, patient and caregiver groups. I gave presentations and recorded responses from the SLaM child and adolescent psychiatry executive group, the Service User Research Enterprise group (SURE), the service user led CRIS Oversight Committee, and SLaM-involved parents, through the BRC patient engagement programme. Because of the focus of one of the projects using the linked data was an investigation into the educational outcomes of children with Autism Spectrum Disorders, I also invited comments on the proposal from the National Autistic Society. I also gained ethical approval from NHS Research Ethics Committee.

7.4.3 Identifying a trusted third party for managing health data linkages

To address the second issue, I provided an overview to the CAG of the advantages and disadvantages of using NHS Digital as a trusted third party to conduct linkages between SLaM and NPD data. I acknowledged that using NHS Digital would not require SLaM to release patient identifiers of over 35,500 names and addresses to the DfE. However, I described this advantage as fairly limited. I argued that the method proposed would involve no release of clinical data to the DfE, and that mental health status data were already collected and available to informaticians working in DfE National Pupil Database Team under their Special Education Need fields. In addition, I explained that DfE informaticians were already contracted to work with highly sensitive information at an individual level (for example, child protection status, benefit status of parents etc.) under comparable data governance standards expected of NHS Digital informaticians, as detailed by HMG Security Policy Framework v10 2013 (SPF).[273] I acknowledged that an additional potential benefit to using NHS Digital was that patient identifiers would be retained within a NHS environment. But after I invited Department of Health (DoH) and DfE to discuss Information Governance standards between their respective departments (in this case HSCIC and DfE Data Division) they advised, and the data controllers accepted, that there was little difference in data security policy. The DoH official responsible for NHS Digital Information Security and Risk Management Policy liaised with the DfE Departmental Security Unit Information Assurance Policy & Governance Team Leader, and reviewed the DfE Data and Statistics Division internal data processing, information handling

controls, and assurance regimes. DoH confirmed that the DfE were in line with government standards and meet equivalent to IG expectations for NHS care system organisations.[278]

To provide further argument for not using NHS Digital as the trusted third party in this linkage, I described two alternatives routes, where NHS Digital performed the linkage and avoided transfer of NHS identifiers to the DfE. One route involved NHS Digital receiving all 15 million identifiers from the DfE, conducting the complex matching with the SLaM identifiers, completing the anonymisation process, and then providing a pseudo-anonymised dataset to SLaM. The second route involved NHS Digital receiving 15 million identifiers from the DfE, conducting the matching process, sending SLaM the controls and cases table with matched SLAM & NPD pseudonyms, and then sending controls and cases with just NPD pseudonym (the DfE remain blinded to SLAM case status) back to the DfE. After this, the DfE would then have to match the education variables of interest on the NPD pseudonym to create a pseudo-anonymised NPD variables table, and finally, send the pseudo-anonymised NPD variables to the CDLS for later matching with CRIS data. I explained that data controllers would likely be concerned with the number of identifiers that would be transferred in both these processes, and the need for sensitive educational variables to be conveyed twice between the parties (DfE to NHS Digital, NHS Digital to CDLS). In addition, and for both options, DfE would need to supply identifiers for 15 million individuals to NHS Digital, which may have contained a number of different addresses for each individual, and then separately convey over 500 education variables per individual, linked by pseudonym to the identifiers. After consulting with the DfE and SLaM data controllers, both expressed concerned that the harm caused to individuals if a breach of data security occurred in either of these processes could be significant, especially given the scale and sensitivity of the educational data, and the very large number of individuals involved. Hence, I advised the HRA CAG that both data controllers preferred to pursue a simpler linkage method, using the DfE to undertake the linkage of identifiers, within their secure environment and with appropriate governance controls using the minimum number of identifiers required.

7.4.4 Equivalence in data security requirements between health and education systems

This third issue was largely addressed by demonstrating data security equivalence between the DfE and DoH standards in processing and storing the data. In the re-submission to the CAG I confirmed that all personal identifiers were destroyed immediately after linkage and validation

by the DfE, and that data was to be anonymised and only analysed within the same secure environment. The table linking NPD and CRIS pseudonyms, would be destroyed after 60 days of the CDLS receiving the data, to permit some additional data cleaning and validation checks. With these additional details, the application was re-submitted and approved (ref CAG 9-08(a)/2013 0048).[274]

7.4.5 Completing the Memorandum of Understanding between Data Controllers

It took some time to formalise a Memorandum of Understanding (MoU) between the DfE and SLaM. This was due to it being the first time an NHS trust in England had entered into a data sharing contract with the DfE, and the lawyers representing both parties took time to become familiar with the legal basis for sharing data in the proposed manner. After a year under legal review, a signed agreement was eventually completed. One of the areas of contention regarded cross-indemnity. Standard legal advice for commercial data sharing often stipulate that each party should indemnify, and keep indemnified the other party, against any claims brought against them despite the proper performance of the Data Activities as envisaged by the MoU. So, taking this linkage project as an example, if someone were to legally challenge SLaM for data that related to the DfE, which they held temporarily during the matching process, then SLaM would honour an agreement to respond the challenge, and vice versa with the DfE. However, if responsibility was shared between parties, it could have potentially created problems in terms of interpretation, especially in relation to data protection compliance, especially for tasks that are time sensitive such as responding to subject access requests. We eventually reached an agreement that the parties would self–indemnify. This decision was aided by the data flows which provided a clear demarcation between DfE and SLaM data systems and procedures, which we came to understand was important when undertaking data processes on behalf of the other data controller. As SLaM and DfE responsibilities for the project were well defined, both agreed that if one party failed in its obligations, it was most likely that enforcement action would be carried out against the party that was in breach of their agreed obligations at that point in the linkage process.

7.5 RESULTS 2

7.5.1 Linkage rates, bias and the impact on education outcome analyses

The overall matching process against any National Pupil Database attendance records provide 29,278 CAMHS-NPD linked records representing a linkage rate of 82.5%. The proportions linked according to DfE matching stages described above: stage 1 - 60.4%; stage 2 - 4.4%; Stage 3 – 1.1% and Stage 4 – 20.7%.

Table 7.2 provides the socio-demographic, clinical and administrative record characteristics of the linked and non-linked SLaM CAMHS sample to NPD data. An odds ratio greater than 1 denotes greater chance of successful linkage compared to the reference. In the adjusted model, we found significant differences in most socio-demographic, clinical and administrative factors. Compared to school age children aged under 7, I found children first referred to CAMHS in late adolescence were significantly less likely to be matched to the NPD absence data, whilst children aged between 7 and 12, were more likely to be successfully matched. Relative to children of White ethnicity, I found other ethnic groups including Asian, Black African and Mixed groups were less likely to be matched. There were no significant differences in successful linkage between children in the lowest and highest quartiles of deprivation, but there was significantly reduced linkage success for children living in neighbourhoods in the 2nd and 3rd quartiles. Analyses of the administrative characteristics show that the post codes which were extracted from clinical episodes of care and that didn't overlap with January census data (i.e. post codes 2,4 and 5) were less likely to link even after adjustment for other potential explanatory variables (see table 7.2). Postcodes which corresponded with a patient being referred to CAMHS after their 16th birthday were also less likely to link compared to those referred when aged under 16.

Table 7.3 provides the socio-demographic, clinical and administrative record characteristics for children seen SLaM CAMHS and their persistent absence outcomes. The adjusted analyses show that presence of an ICD-10 mental health disorder, age at first referral to CAMHS and Mixed ethnic group (relative to white ethnic groups), were associated with an increased risk of persistent school absence, whilst Asian, Black African, Black Caribbean ethnicity, increased neighbourhood affluence was associated with a decreased risk of persistent absence. These effects persisted after both statistical techniques i) using inverse probability weighting, and ii) adjustment for matching probability were applied to reduce matching bias in the adjusted models.

Table 7.2 Socio-demographic characteristics of the Child and Adolescent Mental Health sample linked and non-linked to the national pupil database absence data

	Linked pairs (n=29,278)	Non-linked residuals (n=6,231)	O.R (95% C.I.) for +ve linkage	aO.R (95% C.I.)
Male	16,430 (56.1%)	3,296 (52.9)	*Reference*	*Reference*
Female	12,848 (43.9%)	2,935 (47.1)	0.88 (0.83-0.93)**	1.04 (0.97-1.11)
Age at first referral to mental health services				
Infant (<7yrs)	3657 (12.5%)	535 (8.7%)	*Reference*	*Reference*
Primary (7-12 yrs)	10,980 (37.5%)	1,284 (20.3%)	1.25 (1.12-1.39)**	1.23 (1.10-1.38)**
Secondary (13-15 yrs)	7,048 (24.1%)	1,140 (18.4%)	0.90 (0.81-1.01)	0.98 (0.88-1.10)
College (16-18)	7570 (25.9%)	3228 (52.2)	0.34 (0.31-0.38)**	0.67 (0.59-0.75)**
Ethnicity				
White / White-British	13,838 (47.3%)	2,786 (44.7)	*Reference*	*Reference*
Asian / Asian-British	984 (3.4%)	312 (5.0%)	0.63 (0.56-0.76)**	0.65 (0.56-0.75)**
Black British / African	5,667 (19.4%)	1,181(19.0%)	0.96 (0.89-1.04)	0.82 (0.76-0.89)**
Black British / Afro-Caribbean	1,474 (5.0%)	232 (3.7%)	1.28 (1.11-1.48)**	0.98 (0.84-1.14)
Mixed / Multiple ethnic	2,184 (7.5%)	315 (5.1%)	1.40 (1.23-1.58)**	1.12 (0.99-1.28)
Other ethnic group	1,109 (3.8%)	419 (6.7%)	0.53 (0.47-0.60)**	0.55 (0.48-0.63)**
Not stated	4,022 (13.7%)	986 (15.8%)	0.82 (0.76-0.89)**	0.93 (0.85-1.02)
Resident within Local catchment area	22,481 (76.8%)	4,192 (67.2%)	1.61 (1.52-1.71)**	1.04 (0.97-1.12)
National quartiles of Neighbourhood deprivation				
1st (Most deprived)	14,398 (49.2%)	2,822 (45.3%)	*Reference*	*Reference*
2nd	9,796 (33.5%)	2,179 (34.9%)	0.88 (0.83-0.94)**	0.90 (0.83-0.96)**
3rd	2,956 (10.1%)	762 (12.2%)	0.76 (0.69-0.83)**	0.81 (0.74-0.89)**
4th (Least Deprived)	2,126 (7.3%)	468 (7.5%)	0.89 (0.79-0.99)*	1.03 (0.92-1.15)
Address data available2				
Postcode 1	17,587 (60.1%)	1,987 (31.9%)	*Reference*	*Reference*
Postcode 2	2,956 (10.1%)	990 (15.9%)	0.34 (0.31-0.37)**	0.50 (0.45-0.56)**
Postcode 3	5,776 (19.7%)	1,187 (19.1%)	0.55 (0.51-0.59)**	0.63 (0.58-0.68)**
Postcode 4	1,933 (6.6%)	1,010 (16.2%)	0.22 (0.20-0.23)**	0.35 (0.31-0.39)**
Postcode 5	1,026 (3.5%)	1,057 (17.0%)	0.11 (0.09-0.12)**	0.15 (0.14-0.17)**
Any ICD-10 Disorder	17,749 (60.6%)	3,290 (52.8%)	1.38 (1.30-1.45)**	1.11 (1.04-1.18)**

*P < 0.05,** P <0.01

[1]adjusted for all other co-variates listed in the table.

[2] Post code. For a large proportion of cases there are several addresses available for each case. Therefore, I extracted postcodes according to a hierarchy (Postcode 1 being the highest) which I believed to be most likely to have been the place of residence on the day of the 16th Jan 20XX (variable date) census. [See Figure 7.3 legend]

Table 7.3: Socio-demographic and odds ratios for persistent (>80%) school absence in 29, 278 children and adolescents referred to mental health services

	No persistence Absence (n=23,241)	Persistent School Absence (n=5,635)	O.R (95% C.I.)	aO.R[1] (95% C.I.)	Weighted aOR[2]	Match probability adjusted aOR[3]
Any ICD-10 Disorder	14,004 (60.2%)	3,594 (63.7%)	1.16 (1.09-1.23)**	1.13 (1.07-1.22)**	1.13 (1.07-1.22)**	1.10 (1.03-1.19)**
Age at first referral to mental health services						
<7yrs)	3,031 (13.0%)	298 (5.3%)	Reference	Reference	Reference	Reference
7-12 yrs	9,405 (40.5%)	1,540 (27.3%)	1.67 (1.46-1.90)**	1.67 (1.46-1.90)**	1.67 (1.47-1.91)**	1.60 (1.49-1.84)**
13-15 yrs	5,205 (22.4%)	1,830 (32.5%)	3.58 (3.14-4.07)**	3.65 (3.20-4.18)**	3.71 (3.24-4.23)**	3.66 (3.21-4.18)**
16-18 years	5,600 (24.1%)	1,967 (34.9)	3.57 (3.13-4.06)**	4.20 (3.63-4.86)**	4.15 (3.57-4.81)**	4.70 (3.82-5.78)**
Female	10,023 (43.1%)	2,695 (47.8%)	1.20 (1.14-1.28)**	0.97 (0.91-1.03)	0.97 (0.92-1.04)	0.96 (0.91-1.03)
Ethnicity						
White / White-British	10,651(45.8%)	3,011(53.4%)	Reference	Reference	Reference	Reference
Asian / Asian-British	815 (3.5%)	159 (2.8%)	0.69 (0.58-0.82)**	0.68 (0.57-0.81)**	0.69 (0.58-0.83)**	0.76 (0.60-0.96)*
Black British / African	4,737 (20.4%)	849 (15.1%)	0.63 (0.58-0.69)**	0.68 (0.62-0.74)**	0.69 (0.63-0.75)**	0.71 (0.64-0.79)**
Black British / Afro-Caribbean	1,213 (5.2%)	248 (4.4%)	0.72 (0.63-0.83)**	0.81 (0.70-0.94)**	0.81 (0.70-0.94)**	0.82 (0.70-0.94)**
Mixed / Multiple ethnic	1,653 (7.1%)	483 (8.6%)	1.03 (0.93-1.15)	1.14 (1.02-1.28)*	1.15 (1.03-1.29)*	1.11 (0.99-1.26)
Other ethnic group	905 (3.9%)	195 (3.5%)	0.76 (0.64-0.89)**	0.78 (0.66-0.92)**	0.80 (0.67-0.96)**	0.92 (0.69-1.22)
Not stated	3,286 (14.1%)	694 (17.4%)	0.74 (0.68-0.82)**	0.78 (0.71-0.86)**	0.79 (0.72-0.87)**	0.79 (0.72-0.87)**
Resident within Local catchment area	18,100 (77.8%)	4,064 (72.1%)	0.74 (0.69-0.76)**	0.88 (0.82-0.95)**	0.89 (0.83-0.96)**	0.87 (0.80-0.94)**
National quartiles of Neighbourhood deprivation						
1st (Most deprived)	11,326 (79.7%)	2,884(51.1%)	Reference	Reference	Reference	Reference
2nd	7,891 (33.9%)	1,785(31.7%)	0.89(0.83-0.94)**	0.83(0.76-0.89)**	0.82 (0.77-0.88)**	0.85 (0.79-0.92)**
3rd	2,349(10.1%)	557(9.9%)	0.93(0.84-1.03)	0.74(0.69-0.83)**	0.74 (0.66-0.83)**	0.78 (0.69-0.89)**
4th (Least Deprived)	1,692(7.3%)	413(7.3%)	0.96(0.85-1.07)	0.70(0.62-0.80)**	0.70 (0.62-0.80)**	0.69 (0.62-0.78)**
Address data available[4]						
Postcode 1	14,119(60.7%)	3,170(56.2%)	Reference	Reference	Reference	Reference
Postcode 2	2,287(9.8%)	669(11.9%)	1.30 (1.18-1.43)**	0.71(0.63-0.78)**	0.71 (0.64-0.81)**	0.85 (0.65-1.11)
Postcode 3	4,618 (19.9%)	1,077(19.1%)	1.03(0.96-1.12)**	0.92(0.84-0.99)*	0.92 (0.85-1.00)	1.01 (0.87-1.19)
Postcode 4	1,448(6.2%)	485(8.6%)	1.49(1.33-1.67)**	0.81(0.71-0.93)**	0.82 (0.72-0.95)**	1.14 (0.71-1.81)
Postcode 5	788(3.4%)	238(4.2%)	1.34(1.16-1.56)**	0.93(0.79-1.10)	0.93 (0.78-1.09)	1.85 (0.74-4.66)

*P<0.05,**P<0.01, [1]adjusted for all other co-variates listed in the table. [2] adjusted model with inverse probability weighting for matching included, [3]adjusted model with addition of matching probability estimates entered as a co-variate, [4] See Figure 7.3 legend

7.6 DISCUSSION

Using deterministic and fuzzy matching techniques provided by the DfE, a large-scale dataset was built between NHS child and mental health data and national school administrative data, providing a linkage for 29,278 patients (82.5% of the NHS cohort) to their educational records. Using these data, we found any child or adolescent with a ICD-10 mental disorder had approximately 10% greater likelihood of having persistent school absence, when compared to clinically referred children not meeting threshold for diagnosis. Although there were significant differences in the socio-demographic and clinical characteristics between linked and un-linked NHS samples, effects did not change significantly after matching probability adjustment. This suggests that these effects on were not driven by selection bias from matching errors.

The results suggest, that the approach used in this study can potentially improve the inclusion of socially disadvantaged and vulnerable groups above conventional survey designs.[44] For example, we found no significant differences in data linkages between children in the lowest and highest quartiles of deprivation. This study demonstrates how routinely collected NHS data and non-NHS administrative sources can be linked without individual consent in England.

Overall, I found that only 17.5% of the clinical population were not successfully matched. Whilst enrolment at a non-state maintained school or independent school may explain a proportion,[270] a significant minority were not matched due to administrative factors, which may include missingness or inconsistencies of the matching identifiers, as demonstrated by the effect of post code variation in the analysis, or errors secondary to the matching process. There have been very few studies conducted which examine linkage errors in children, especially where the non-linked group are not subject to consent related bias. However, my study findings showing significant differences in sociodemographic factors and differential linkage rates , especially between white and ethnic minority groups, have been found in a number of studies. [262,266,279] I was unable to examine what clerical or patient factors may be driving these increased errors. However, previous studies have suggested that ethnic minorities are more likely to have misspelt names, inaccurately recorded dates of births, and higher levels of residential instability, which may be applicable to this sample.[266,279] I found certain age groups, particularly the 7 to 12 year old category, were associated with a greater likelihood of linkage. This may be due to the greater availability of accurate personal identifiers in the

record as this group, as their potential exposure to CAMHS services whilst at school will be longer than other age groups. Similarly, having a ICD-10 mental disorder, which also had an increased likelihood of linking with the school data, may be related to identifier accuracy, as their higher levels of psychopathology will be associated with greater clinical contact, and potentially higher clerical accuracy in recording personal identifiers. It also more probable that those higher levels of psychopathology will have longer durations of care that overlap with the school census date.

I found a U-shaped distribution in neighbourhood deprivation and likelihood of linkage. Compared to areas with the highest deprivation, areas within the 2nd and 3rd quartiles showed significantly reduced likelihood of linkage, but the most affluent areas showed minimal difference. This could relate to families from affluent areas being able to comply with the administrative process, and/or correct administrative errors, and families from the highest deprived areas having greater need and hence higher clinical contact with services. Both these factors may improve clerical accuracy and concordance with school data. Families from 2nd and 3rd quartiles may have less of both these characteristics, reduce their likelihood of linkage. The current data available in this study does not permit this hypothesis to be tested, but findings suggest that a more detailed extraction examining frequency of clinical contact with services and data linkage outcome is an area for future work.

The findings show that potentially 17.5% of the clinical population were not matched. A considerable proportion may have been missed due to technical errors in the linkage approaches. This provides an argument for government departments to trial more modern approaches to data-linkage, for example using probabilistic methods (as described in chapter 2).[60] The largely deterministic matching process, although potentially providing high precision, is likely to have contributed to missed match rate (i.e. false negative non-matches). Child names of foreign origin (i.e. not commonly associated with the predominant ethnic group population) are more likely to be inaccurately entered into administrative systems.[266] Hence, the deterministic process which offers little flexibility in matching misspelt names may be a reason why ethnic variation may contribute to false negatives.

The study provides an example of how potential non-random loss between routinely collected health and non-health linked data can be adjusted by weighting techniques. Differential linkage error by ethnicity, social disadvantage and clinical factors can introduce significant selection bias, leading to inaccurate risk factor-outcome estimates, which in turn may have significant

impact on the validity of the research findings using the linked data. I was only able to determine that linkage error did not lead systematic biases, and provide misleading positive estimates between ICD-10 mental disorder and persistent school absence, because I had source information available data to examine missed linkages. Without this information, potential linkage error could be introduced, and I would not be aware of whether there was need for it to be accounted for in subsequent analyses.

In this study, the demonstration of matching probability adjustment and inverse probability weighting was intended to illustrate how linkage bias may be reduced, not as a definitive analysis of these data. Future work examining on how improving linkage techniques, coupled with newer methods for handling uncertainty in analysis of linked data, should help improve the generalisability and quality of future population based linkage studies.

7.6.1 Limitations of the matching methods and matching evaluation

This study has a number of limitations. I was unable to assess false positive matching, nor able to assess risks for the lower confidence matching (DfE stages 2-4, described above), and the potential effects on school outcome analyses. No shared unique identifier exists between NHS and educational services, nor were their governance arrangements or sufficient resources in place to manually compile a NPD-NHS linked gold-standard data. Another limitation of the matching methodology is the limited number of address identifiers that could be used. For example, due to governance constraints I was unable to use first line of the address, which again limited the capacity to potentially check for coding errors in the postcode. Another contributing factor to linkage error was the age of the child. A substantial number of adolescents were seen in CAMHS aged 16 and 17 years, and would not have data on the NPD if they were no longer attending school. Similarly, I was unable to determine who was not eligible for matching due to complete private or home school educational provision which may be around 5% of the sample (see chapter 6).

7.6.2 Applying existing legal and ethical frameworks to data linkage between health and education data

Before embarking on the linkage project between NPD and CRIS CAMHS data, I was aware that repurposing routinely collected individual level data for research, which involved child mental health and education service use, was potentially controversial, especially as I was not actively seeking consent. Also, the project development began at time when Care.Data looked close to being disbanded.[280] In addition, significant changes to EU data protection law were being proposed, which if not amended, would have prevented the data linkage methods being applied.[281] I was concerned therefore that the public could have significant concerns about the linkage process, especially in relation to vulnerable children, who may be particularly harmed by exposure to stigma and the loss of trust in care services following a breach of privacy. I knew measures such as anonymisation did not solve all ethical, legal and technical problems.[74] Hence, I expected that higher thresholds for ensuring social benefit from data linkages may be imposed by data controllers and custodians in order to preserve public trust.

7.6.3 Implementation challenges to the data linkage between health and education data

Some lessons I learnt during study, may be of use to future health orientated linkage projects where individual level consent is not available or practicable to obtain. The first element was what might be called establishing *the social license* [74] to use personal health and education data for data linkage and research. This activity included articulating a clear purpose for the linkage which was recognized as beneficial by the public or those potentially involved as data subjects, and that the potential risks to individuals or public institutions were tolerable in relation to these benefits. Without the evidence of the proposal being scrutinized and ultimately accepted by those potentially involved as data subjects, and the public institutions/services who act as controllers of the data, it would have been difficult to sustain a case for public benefit – in fact this was one of the reasons why my first application was not approved by the HRA CAG. To prove I had *social licence* to conduct the linkage work, I gathered supportive evidence from a number of sources including service users, clinicians, academics, advocacy groups, governance leads; all, who I viewed, may have had stake in the process and outcomes of the data linkage project.

The second lesson related to fulfilling the professional mandate for properly conducting the linkage process and related research activity. This involved making sure the proposal complied with the known legal, technical and ethical frameworks that governed health data use, and any additional safeguards deemed important by the data controllers and custodians. The technical aspects were not just confined to data security, but also involved preparing the data to ensure the most accurate match, to reduce error and redundancy in later analysis. Fulfilling the mandate also involved the creation of formal contract between the parties involved in controlling, sharing, processing and using the data. This mandate committed us to conduct appropriate analysis and dissemination of the linkage related research, so that I could sustain the social license for future research activity. This may be especially pertinent in England as linkage driven research of routinely collected public service activity is in its infancy, and benefits are yet to be comprehensively established.

Another element, which is difficult define,[282] was the need to establish trusted relationships between all the parties during the process of steps 1 and 2. These data linkage projects involved public service organisations, and the government funded committee's acting on behalf on the public, taking on additional risk in approving and conducting the linkage process. In the case of the DfE and SLaM data controllers and custodians, the potential risk of harms from the data linkage process were small, but nonetheless could be viewed as an unnecessary addition to the everyday operational risks they normally managed. For both data controllers, it was not a core part of their business to ensure these data linkages were conducted. However, at the early stages of the project I was fortunate that several individuals representing both data controlling organisations were able to quickly establish and sustain a mutual interest in the data linkage's purpose, a tolerance of the potential risks, and the capability and authority to help progress the project when it became stuck. Furthermore, these individuals have remained involved in the project from its inception in November 2011, linkage completion in August 2015, up to the ongoing analysis and dissemination of findings. Over this period, they remained accessible to one another and keen to maintain an open dialog.

Given the time and resources spent to set up this linked data resource, and the potential it holds, it is important that these resources are maintained, and remain accessible for re-use in the future. Without developing specific data sharing agreements between the parties, it can be difficult to establish a collaborative relationship with good governance structures between the controllers, linkers and analysts. Without these structures, there may be a tendency for data controllers to agree to link data only via a 'create and destroy' approach. When these

agreements are made, the linked data and any interim datasets are destroyed at the end of each project.[283] I believe this maybe unethical in terms of waste and scientifically unsound as prior analyses cannot be re-examined. It also re-exposures data subjects to the potential risks of sharing personal identifiable information again across different agencies should the linkage need to be repeated in the future.

The implications of a trusted and collaborative relationship between data controllers, data processors (i.e. data linkers) and analysts are key, not just to the successful completion and continuation of the project, but the quality of the research output derived from the data. Trust in the governance structures between these parties enables some flexibility with the 'data separation principle.' This principle describes a common practice for data linkage research, where identifiers (e.g. names or date of birth) are kept separate from attributes (in this case health or education data), to protect privacy and avoid disclosure during the linkage process.[265] While the separation principle might reduce the risk of identification, it can increase the risk of biased analyses.[14] In order to reduce these biases, governance arrangement between linkers and analysts should permit information to be shared on which groups are disproportionately affected by linkage error. Doing this can then enable discussions on how linkage errors can be mitigated, either through changing the linkage process or modifying the analyses. Through demonstrating the potential bias incurred through non-linkage, the study in this chapter supports the argument that data providers who wish to build linked resources to analyse routinely collected data, need to provide linked and unlinked records in order to take account of biases. I hope in further linkage work between SLaM and DfE, I will acquire further detail on DfE's linkage uncertainty (i.e. metric's which quantify the probability of each linked record being a false-positive matches) which can also be included in later outcome analyses to mitigate potential errors incurred via the linkage process.[284]

Finally, ensuring that sufficient resources were available to see the project through to completion was essential. This required a sustained commitment from a number of people and within SLaM BRC and the DfE. Both institutions have facilities, and teams within them, equipped to manage the hosting, secure access and development requirements to link complex clinical and social data resources. The CRIS CAMHS linkage to the NPD project was supported by the shared expertise of academics, project managers, service users, clinicians, health and education informaticians and NHS & DfE clinical governance leads and legal teams. Nonetheless, the CRIS CAMHS linkage to the NPD took over 3 years to complete. Although not exclusively devoted to the linkage project, both SLaM BRC and the DfE had three whole

time equivalent staff who provided managerial, administrative and analyses support to CRIS and NPD throughout the project. I believe these elements helped to develop and sustain trust over the long duration of establishing the approval to conduct this data linkage project.

7.6.4 Conclusions

In this chapter, I provide an example of how data linkage projects can be completed using routinely collected NHS and non-NHS resources. Data linkages methodologies hold significant opportunities for public services research and policy development including child mental health services research, and can be highly efficient relative to other epidemiological approaches. The regulatory and technical issues for data sharing between health and non-health services are challenging in England. Certainly, to develop and improve linked data resources, partnerships between academic and government institutions should continue to explore public opinion and develop guidance on building a social license for the sustained use of linked data.

Record linkages are a valuable enhancement to child-based longitudinal studies and clinical registries, allowing evaluation of questions relevant to public health and social care policy. The study results suggest, that opt out consent approaches may improve representation of more socially disadvantaged populations.[44] Nevertheless, whether using opt in and opt out consent process, possible biases due to linkage error can be important and need to be addressed when analysing and interpreting results.

I hope this account may provide a useful guide for other health and educational services wishing to build information resources using linked administrative data. In time, I hope these resources will generate a wider network of fine-grained data and analytical expertise, which can be used for research to inform commissioning and service provision and better meet child mental health needs within the population.

CHAPTER 8. AUTISM SPECTRUM DISORDERS AND RISK OF SELF-HARM IN ADOLESCENCE: A RETROSPECTIVE COHORT STUDY OF 113,545 YOUNG PEOPLE IN THE UK

8.1 SUMMARY

Background: Presentation to emergency care with injuries related to self-harm is one of the strongest predictive factors for later suicide attempt, with 26% of future suicide attempts attributed to self-harm in adolescence and young adulthood. Recent findings show individuals with autism spectrum disorders (ASD) have a 2-3 fold increase risk of premature mortality compared to the general population, with suicide as the leading cause. The risk of self-harm severe enough to warrant emergency treatment has yet to be robustly evaluated in ASD. This study assessed whether individuals with ASD are at increased risk of self-harming in adolescence.

Method: I conducted a population based retrospective cohort study. The source population were residents of four South London boroughs, aged 11-17, attending secondary school identified from the NPD. Exposure data on ASD status were derived from the pupil database. Outcome (self-harm data) were derived by linking the education record with CAMHS records.

Results: Among 113,543 adolescents attending secondary school, 186 boys (0.3%) and 834 (1.4%) girls presented to accident and emergency with self-harm; less than 50% of whom were previously known to NHS mental health services. In the sample, 2463 adolescents were identified with ASD. For boys, there was a significantly increased risk of self-harm associated with ASD (aH.R 2.79, 95% confidence interval 1.47 to 5.09, P<0.01) after adjustment for potential confounding factors, including baseline behavioural & emotional problems, academic attainment, persistent school absence, exclusion, socio-economic status, being in local authority care, and hyperkinetic disorder diagnosis. For girls, ASD was not associated with elevated risk, but a number of educational, social and clinical related factors were identified as significant predictors of self-harm, including persistent school absence (aH.R 2.84, C.I 2.70-3.51, P<0.01), and being in higher quintiles of academic attainment (aH.R 1.35, C.I 1.04-1.77, P=0.03).

Conclusions: This study provides robust evidence that ASD, and a number of other educational factors, are population level risk factor for self-harm. Risk is not equal across gender, with ASD associated a with greater susceptibility to self-harm only amongst boys. These finding are an important first step in developing early recognition and future prevention programmes within schools and other child orientated services.

8.2 INTRODUCTION

Autism Spectrum Disorders (ASD) are childhood-onset neurodevelopmental conditions, characterised by a 'spectrum' of social and cognitive impairments. Over the last four decades, the recognition of childhood ASD in the population has increased exponentially.[285] The widening of diagnostic criteria and increased detection has moved ASD from a rare neurodisability managed in specialist clinics, to becoming a public health concern affecting around 1% of the population.[286,287] Two broad questions predominate ASD public health service research. The first relates to how early identification and support to families of children with ASD can improve childhood social development and function.[288–290] The second question concerns how services can reduce the health and social disadvantages which affect individuals with ASD over the life course. Individuals with ASD carry a 2 to 3 fold greater risk of premature mortality notwithstanding the developmental comorbidities which tend to co-occur with ASD.[291,292] Over 50% of adolescents with ASD fail to complete higher education or find employment [293,294] and most will remain heavily dependent on their family throughout adulthood.[295–297] Both questions need equal attention. Whilst there is some evidence that quick detection and support in early childhood may enable important gains in social capabilities, it appears unlikely that it will mitigate all risks for later impairment, especially psychiatric morbidity which has a considerable impact on ASD function.[298,299]

The psychiatric burden and related impairment carried by children with ASD is sizeable. Over 70% of children and adolescents with ASD will develop a least one psychiatric disorder.[118,286,300] Despite the high prevalence of psychiatric impairment, children and adolescents with ASD are likely to face greater difficulties in getting their psychiatric conditions recognised. Communication difficulties combined with our current methods of detecting psychiatric problems, make it harder for individuals with ASD with psychiatric morbidity to attract help, especially for the more internalising conditions such as anxiety and depressive disorders.[118] This is further compounded by adolescents with ASD being at greater risk of being social marginalised [301] and having less access to effective treatment, even once their difficulties are recognised.[302,303] These factors may accumulate in young people with ASD, placing them at greater risk of adverse consequences from their psychiatric disorders, including severe self-harm.

Self-harm presentations to hospital represent one of the strongest risk factors for future suicide attempt,[304] increasing the risk approximately 10-fold compared to the general population.[305] Self-harm is defined by the National Institute for Health and Care Excellence (NICE) as, "any act of self-poisoning or self-injury carried out by an individual regardless of motivation.[306] It is common among adolescents. A large scale systematic review reported 13% of adolescents have self-harmed at some point in their childhood.[307] A recent meta-analysis showed that previous self-harm, or thoughts of self-harm, identified adolescents and young adults who are most vulnerable group for attempted suicide.[308] A subsequent analyses found prior self-harm accounted for 26% of future suicide attempts in adolescents and young adults.[309] Adolescent presentation to hospital occurs in only one in eight self-harming episodes in the community,[246,310,311] generally when injuries are too severe to be self-managed.[312] Within the UK, a study of serious case reviews found 10–20% of young people who die by suicide, visit a hospital for self-harm in the year prior to their death.[243,313]

Incidence of self-harm is concentrated in younger age groups: the majority of cases are in under 35s and with peak age at presentation in women between 15 and 19 years, and in men between 20 and 24.[309,314] Population surveys of adolescents show self-harm prevalence is different between genders, with approximately 11% of girls reporting self-harm in the previous year compared to 3-6% of boys.[315,316] Depression and anxiety, low self-esteem, impulsivity, attention and conduct difficulties are the most replicated risk factors for self-harm.[246,316,317] Marginalised young people including victims of maltreatment, those with lower socio-economic status, school excluded or with prolonged absence from school are also potentially more at risk. [318–321]

Findings emerging from recent epidemiological studies on suicidal behaviour in adulthood certainly support the hypothesis that higher rates of self-harm could be expected in adolescents with ASD. A large-scale population study showed that suicide is a leading cause of premature death in adults with autism.[291] A clinic based study found 66% adults newly diagnosed with ASD, reported that they had contemplated suicide (UK general population prevalence is 17%) and 35% had planned or attempted suicide.[322] Croen et al. found the risk of suicide attempts were fivefold higher in adults with ASD compared to non ASD controls. They also found adults with ASD were significantly less likely to be diagnosed with alcohol abuse/dependency and to self-report alcohol use, suggesting that substance misuse, a strong contributing factor in the general population, may not have the same attributable risk in ASD samples.[323]

As far as I am aware, there have been no prospective cohort studies which have examined the association between ASD and self-harm in adolescence.

In the absence of epidemiological studies, clinical opinions have suggested self-harm may occur less often among adolescents with ASD than in the general population.[324] A potential issue with research that examines self-harm behaviour in ASD is the possible conflation with self-injurious behaviour. Self-injurious behaviours are diverse and often highly repetitive and rhythmic types of behaviours (for example head banging, hair pulling, arm biting, eye poking, and skin scratching), that occur without an apparent intent of wilful self-harm, and result in physical harm.[325,326] In contrast to adolescent self-harm, self-injurious behaviours among young people with ASD are associated with lower chronological age,[326] do not show any particular associations with gender, ethnic background or socio-economic status[104,327] and require a different approach to management.[286,328]

From the limited research conducted, findings show adolescents with ASD are at greater risk for reporting suicidal behaviours [151] - a broad term which captures thoughts, plans, and attempts to end one's life - which in the general population are strongly associated with self-harm.[329] One clinical study found over 1 in 6 young people with autism spectrum disorders (ASD) will contemplate or attempt suicide during childhood, making them 30 times more at risk than typically developing children.[150] However, the clinical implications of these studies are difficult to judge, as qualitatively diverse events have been aggregated into binary outcomes. For example, in one of these studies [151], a child who was rated on one questionnaire item by a caregiver as *sometimes talks about harming or killing themselves* had an equivalent outcome status to another child who was rated *often attempts suicide.* This is problematic as clinical and policy implications would differ if ASD was not associated with increased risk of suicidal ideation, but a greater risk for suicidal attempt. Furthermore the methodological weaknesses, including the cross-sectional designs, small and selective nature of the samples, and lack of adequate adjustment for possible confounding factors or comparable control groups,[150,151,330,331] further limit the interpretation and generalisability of these findings.

To address these issues, I conducted a historical cohort data linkage study using contemporaneous, routinely collected data from school census records matched to psychiatric liaison records. As described in chapter 6, this longitudinal data captures at an individual level, the whole adolescent population continuously resident within four large boroughs within South

London, and provides a very accurate population denominator. Using these data, I aimed to provide age and gender stratified incident rates, and to test whether adolescents with ASD had a greater risk of self-harm presentation compared to those without ASD.

8.3 METHODS

8.3.1 Sample

I used anonymised NPD data comprised of children and adolescents enrolled in state maintained education, and resident within a local catchment of four South London Boroughs (Southwark, Lewisham, Lambeth and Croydon) linked to SLaM CAMHs electronic health records[129,156] (as described in chapters 2, 6 and 7) via the CRIS system.

Using these longitudinal school census data, I identified a dynamic cohort of adolescents (aged 11-17 inclusive) who had resided within the 4 boroughs between 1st January 2009 until their eighteenth birthday or 31st March 2013, whichever was sooner (please see cohort table in Appendix A which provide age, year of study entry and the duration of follow-up). During this period SLaM provided 24-hour psychiatric liaison services within the local catchment's four main acute NHS trust Emergency Departments (ED): St Thomas' Hospital, King's College Hospital, Croydon University Hospital and University Hospital Lewisham, which were all staffed by psychiatric liaison nurses and psychiatrists, and recorded self-harm attendances in the ED using the SLaM electronic health record system (ePJS). All four EDs have policies of referring all attendees with self-harm for a SLaM psychiatric assessment and of recording these referrals regardless of whether individuals wait to be seen.[251]

8.3.2 Measures

Outcome

The primary outcome was first attendance to acute hospital services with self-harm behaviour. To identify cases of self-harm, I used a similar methodology described by Polling et al., and defined self-harm according to the National Institute for Health and Care Excellence (NICE) definition; " any act of self-poisoning or self-injury carried out by an individual regardless of motivation".[332] However to exclude self -injurious behaviour typically associated with non-harming intention, I included the caveat that self-harm needed to be wilful.[333] Presentations were excluded if the details of the self-harm episode were detected co-incidentally during

history taking and had occurred more than seven days prior to ED presentation. Any ingestion of non-recreational drugs above the prescribed dose identified as self-harm by the individual or ED staff was coded as self-poisoning. Use of recreational drugs was coded as self-poisoning where the patient reported intent to self-harm. Episodes were coded as self-injury where any intentionally self-inflicted injury, however superficial, had occurred but not where threats or gestures to self-harm had not resulted in injury. All attempted hanging, jumping from a height and immersion in water with intent to drown was coded as self-harm and categorised as "other" regardless of whether injuries were sustained.

To ascertain the first self-harm event using the above definition, I used CRIS-HES linked data for both Admitted Patient Care and ED episodes of care (see figure 8.1, and chapter 6, figure 6.1).[156] HES data were available within CRIS for all adolescents who had any contact with SLaM services over the observation, and non-SLaM data for all those resident within the local catchment area at the time of their hospital use. Automated data extraction steps were developed to extract the first episode of self-harm, as follows

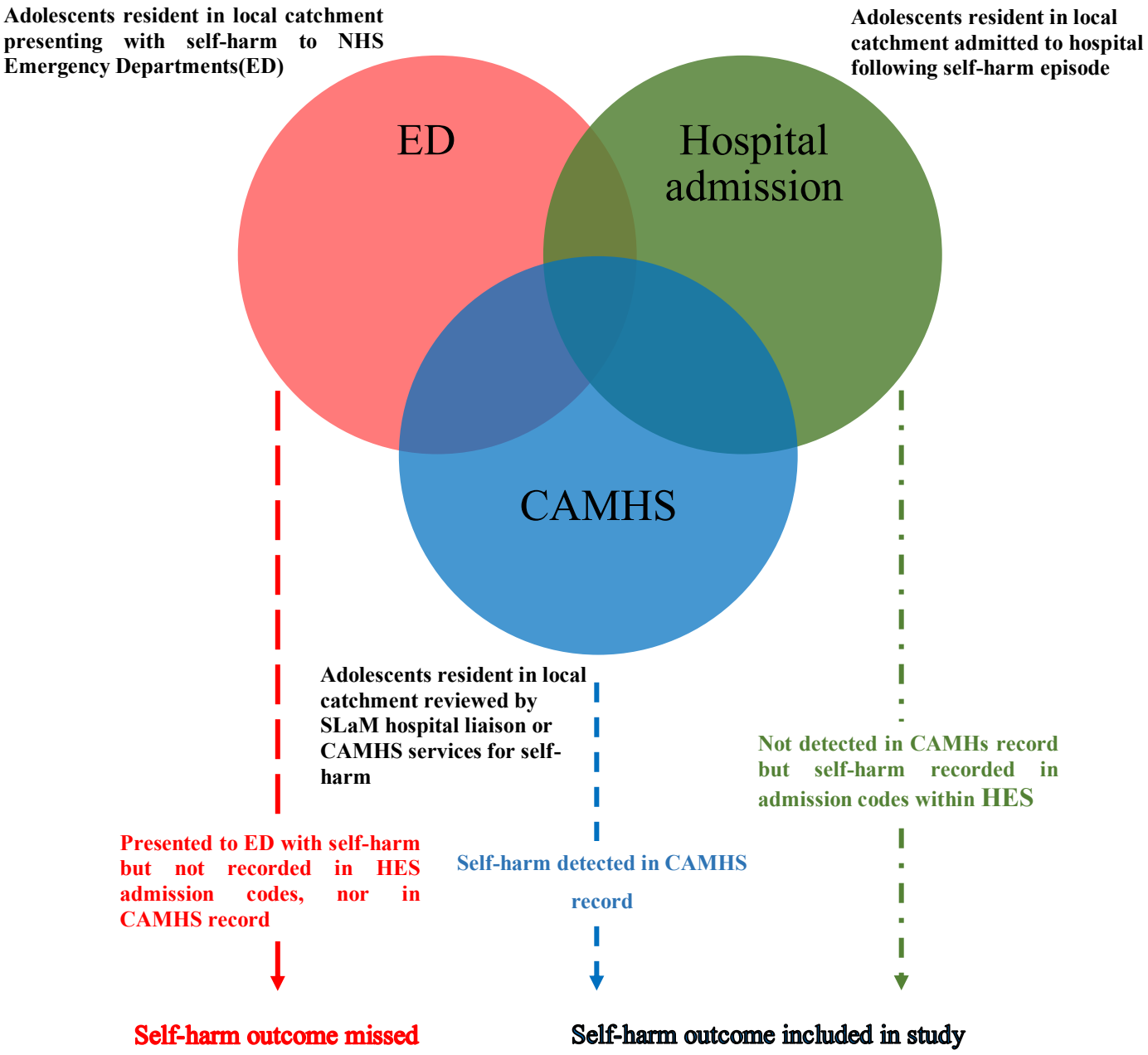
- Step 1. All HES ED or emergency admission from Admitted Patient Care (APC) records were retrieved within the observation window for children aged between 11 and 17 (at the time of the admission) with a home address within the four South London boroughs [An admission was considered to be an emergency if the HES method of admission variable ('admimeth') was 21-24 or 28].

- Step 2. All cases identified in step 1 were retrieved who had any linked CRIS record entered after 12 hours from time and date of ED admission **OR** any cases identified via emergency APC entry with ICD-10 self-harm diagnosis codes (see table 8.1). I considered any multiple admissions within 1 day of each other, or relating to a hospital transfer, to be the same admission.

- Step 3. Any free text records were retrieved with the self-harm key words entered in structured assessment forms, risk proforma, free text note or correspondence item in CRIS.

- Step 4. Documents identified from step 3 were rated by one of two clinical coders, and coded for presence of self-harm, type of self-harm. Previous research had shown this approach has high inter-rater reliability (presence of self-harm kappa 0.85, type of self-harm 0.87).[251]

Figure 8.1 Data sources used to capture first self-harm event from HES administrative database linked via CRIS to SLaM Child and adolescent Mental Health Data



Exposure: Autism Spectrum Disorder

I used the NPD special education needs (SEN) register to identify all ASD diagnoses. In the UK, schools can seek external advice and resources from the local educational support (LEA) services, the local Health Authority or from Social Services when unable to meet the learning needs of an enrolled child. This extra provision, called "school action plus", may include advice from a Speech and Language Therapist, an Occupational Therapist or Specialist paediatric services. It may also include one-to-one support and the involvement of an Educational Psychologist. When these extra provisions are not sufficient, the school may request an LEA assessment under a statement of special education needs (SEN).

Table 8.1 Definitions and International Classification of Diseases (ICD-10) diagnostic codes used to classify emergency admissions for self-injury

Self-harm description	ICD-10 code
Intentional self-poisoning by and exposure to…	
…drugs	X60-X63
…other and unspecified drugs, medicaments and biological substances	X64
…alcohol	X65
…organic solvents and halogenated hydrocarbons and their vapours	X66
…other gases and vapours	X67
…pesticides	X68
…other and unspecified chemicals and noxious substances	X69
Intentional self-harm by…	
…hanging, strangulation and suffocation	X70
…drowning and submersion	X71
…firearm discharge	X72-X74
…explosive material	X75
…smoke, fire and flames, or steam, hot vapours and hot objects	X76-X77

…sharp/blunt objects	X78-X79
...jumping from a high place	X80
…jumping or lying before a moving object, or crashing a motor vehicle	X81-82
…other specified means	X83
…unspecified means	X84
Personal history of self-harm	Z91.5

The SEN statement (now referred to as an education and health care plan) was a legal document issued by the LEA for children who need substantial additional support in school because of learning or behaviour problems. All children who are recognised as being school action plus or have provision under a statement of educational need (approximately 7-9% of children) had reasons registered under a specific category of need,[334] which included the following:

1. Specific Learning Difficulty
2. Moderate Learning Difficulty
3. Severe Learning Difficulty
4. Profound & Multiple Learning Difficulty
5. Behaviour, Emotional & Social Difficulties
6. Speech, Language and Communication Needs
7. Hearing Impairment
8. Visual Impairment
9. Multi-Sensory Impairment
10. Physical Disability
11. Autistic Spectrum Disorder (ASD)
12. Other Difficulty/Disability

The SEN register has been used in epidemiological studies to identify ASD populations, and has high diagnostic specificity.[125,335] The NPD provides up to two different SEN codes (primary and secondary) for each child, to allow for multiple needs to be captured. For the study sample, I identified ASD from either of the primary or secondary SEN fields, and from any of the census periods available between academic years 2004/5 to 2012/13.

Confounders and Risk Factors: Socio-demographic factors

NPD census data from the last available academic year was used to provide pupil characteristic details on gender, ethnicity and English as a second language. This last characteristic is identified by schools via parental report of English not being the primary language spoken within the child's home. Where these socio-demographic data were missing in the NPD census data, I used linked health data to replace the missing values.

For other baseline characteristics, other school factors were collected from the NPD census and other NPD registers, and extracted from the academic year prior to the date of entry into the study. These characteristics included free school meals eligibility (a proxy for low socio-economic status), neighbourhood deprivation, and whether, owing to child protection concerns, the child was under care of the local authority (i.e. a looked after child). I categorised neighbourhood deprivation according to the Index of Multiple Deprivation scores based on residential postcode, with use of quintile cut-off values for England.[135]

Other special education needs

I used the SEN register to identify other special educational needs. For those categories which identified a developmental condition, I used either of the primary or secondary SEN fields, from any of the census periods available between academic years 2004/5 to 2012/13. These included all learning difficulties, hearing, vision or physical disabilities, or special, language or communication categories. To reduce the potential for reverse causality between self-harm and the behaviour, emotion and social problems SEN category (i.e. where self-harm leads to school recognition of this need), I only coded behaviour, emotion and social problems at baseline i.e. using all relevant codes in SEN register data until the academic year of entry into the study.

Educational attainment

I used educational attainment at KS2 examinations also known as Standardised Attainment Tests (SATs). These are taken in year 6 (children aged 10-11) at the end of Primary schools. SATs are taken in three core subjects: English, Maths and Science, and for each subject, a total test mark is generated. I calculated an average mark score from the available results, and created a ranked z-score. This ranked score was then divided in 5 quintiles. The KS2 test marks data will be missing if the KS2 level is "B" (pupils working below test levels); I therefore assigned anyone with a 'B' as within the lowest quintile.

Educational attendance and exclusion.

I created a binary outcome marker of poor attendance for the academic year before they entered study. I categorised pupils as poor attenders if they had recorded less than 80% school attendance for the total number of possible school sessions available since their enrolment for that academic year (one session is equal to half a school day). A binary marker, using historical NPD exclusions data, was also calculated for any child that had a prior record of exclusion (fixed term or permanent) up to the point of study entry.

Hyperkinetic Disorder co-morbidity

Using the linked NPD-CAMHS data, and incorporating the methodology described in previous studies,[129,160] I extracted any ICD-10 recorded comorbid psychiatric diagnoses of hyperkinetic (F90) disorder from CRIS.

Prior attendance to CAMHS services and diagnostic data

In the subset of adolescents who had attended ED with self-harm, I extracted from CRIS any record of previous contact with mental health services, and ICD-10 diagnosis data from any time point within the observation window up until their 18th birthday. ICD-10 Axis one diagnoses were categorised into: substance misuse disorders (F10-F19) psychotic (F20–F29, F31, F32.3, F33.3), depressive disorders (F32), anxiety, stress and emotional disorders (F40–42, F43–F48), eating disorders (F50), childhood-onset emotional and behavioural disorders (F91-F98). Low frequency psychiatric diagnoses were collapsed into a single category labelled "Other". Adolescents without a CRIS diagnosis, or only detected via HES APC codes were coded as "No diagnosis recorded"

8.3.3 Analyses

Previous work has established differential risks for self-harm according to gender, which appears to be particularly marked in adolescents and clinical samples.[336] Hence all analyses were conducted separately for girls and boys. I used the statistical package Stata MP (version 14). Statistical disclosure rules (a condition of using the NPD dataset) required us not to publish counts of less than five, I do not present exact numbers of self-harm for certain groups.

I first derived numbers and proportions of incident self-harm following entry into the study, by each year of age 11-17. The numerator were adolescents who first presented within the age

category with self-harm, the denominator was the population within that age band who were at risk. Confidence intervals for each age band were calculated using the Stata 14 ci command which provides exact (Clopper–Pearson) confidence intervals.[337] Having checked proportional hazards assumptions, Cox regression procedures were used to calculate the unadjusted hazard ratios (and their 95% CIs) for a number of potential risk factors, justified on basis of previous research on their relationship with adolescent self-harm.[246] I then present an adjusted analyses for the hazards of presenting with self-harm, which adjusted for all covariates examined.

I conducted several sensitivity analyses, to aid interpretability and reduce potential biases. The first sensitivity analysis restricted the sample to pupils only attending mainstream secondary schools, excluding those attending special schools or pupil referral units, as the latter schools were likely to have populations with much greater psychiatric morbidity. The second sensitivity analyses restricted the cohort to those who entered at age 11, hence reducing the potential for variation in baseline effects to differ because variable age/maturity at study entry. The third analyses, used a multiple imputation approach to examine whether missing data related to non-matching between the national pupil database and SLaM self-harm data caused substantial changes in direction of size of the effect between ASD and self-harm. Because complete outcome data was available and there were a considerable number of predictor variables related to non-linkage (see chapter 7), I assumed the data was missing at random.[338] I created 10 imputed datasets (m=10), as recent recommendations are to perform at least as many imputations as the proportion of missing cases in a study,[339] and used the distributions from the complete case dataset to cross check against the imputed dataset. The cox analyses were then repeated in the imputed sample.

8.4 RESULTS

Figure 8.2 shows how the adolescent population sample and self-harm cases were ascertained from the National Pupil Database, HES and CAMHS databases. Overall, using residential census data from the National Pupil Database (NPD), 113,288 adolescents were eligible for entry into the study (i.e. aged between 11-17 years and resident within Southwark, Lambeth, Lewisham and Croydon throughout the follow-up period).

Figure 8.2 Sample and self-harm case ascertainment using National Pupil Database, HES and CAMHS databases.

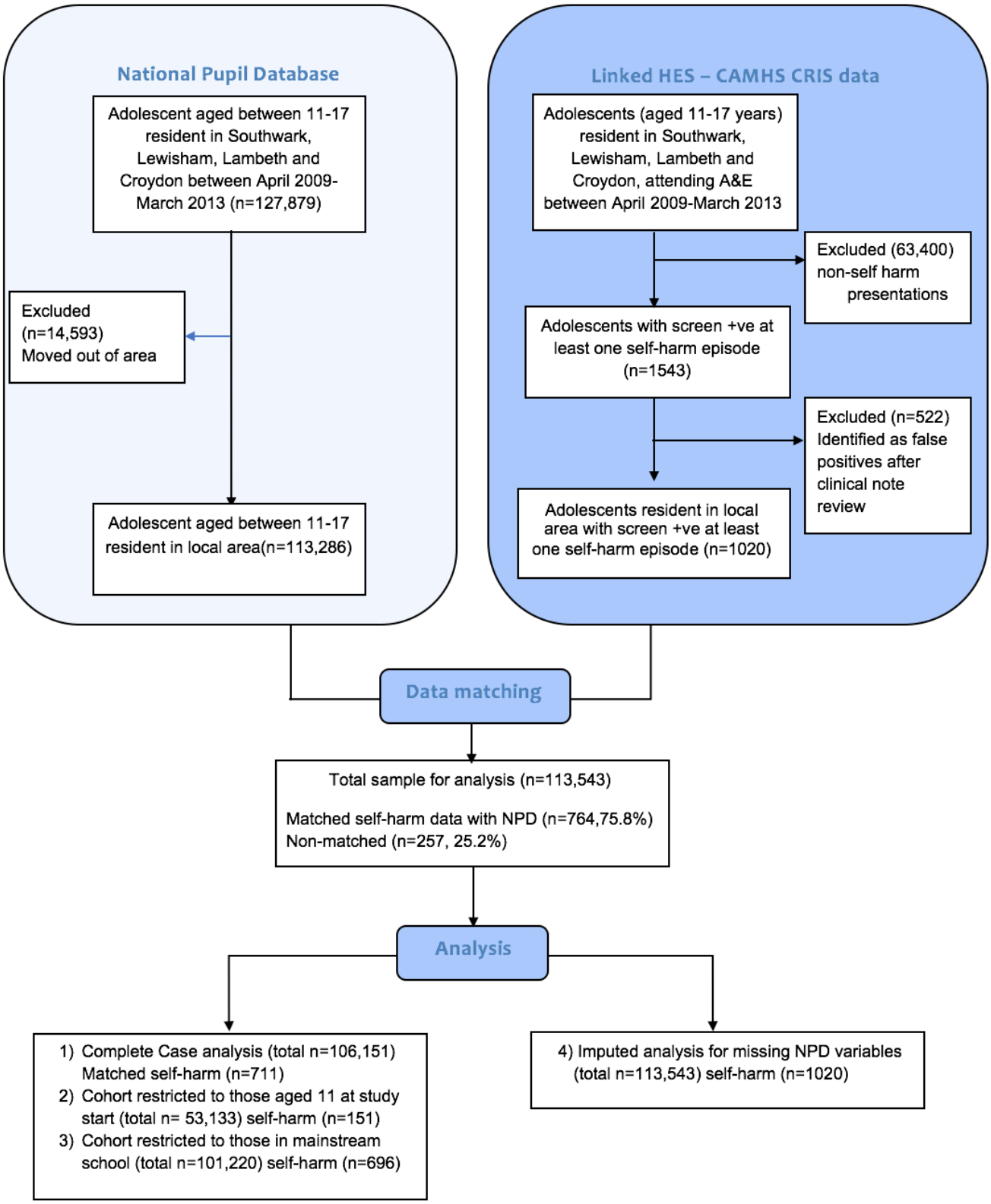


Table 8.2 Cross-sectional characteristics of those adolescents presenting with self-harm by summarised measures of self-harm, and other clinical factors

Characteristics	Self-harm presentations (n, %)	
	Male (n=186)	Female (n=834)
Mean age at first self-harm presentation (SD)	15.9 (1.9)	15.6 (1.4)
Known to MH services prior to self-harm	83 (44.6)	407 (48.8)
Ethnicity		
White	88 (47.3)	357 (42.8)
Black	28 (15.0)	212 (25.4)
Asian	12 (6.5)	47 (5.7)
Mixed	16 (8.6)	102 (12.2)
Other	12 (6.5)	29 (3.5)
not disclosed / unknown	30 (16.1)	87 (10.4)
National neighbourhood deprivation		
Most deprived quintile	63 (33.9)	320 (38.4)
2nd	79 (42.5)	330 (39.5)
3rd	32 (17.2)	120 (14.4)
4th	8 (4.4)	44 (5.3)
Least deprived quintile	4 (2.2)	20 (2.4)
Type of Self-Harm		
Self-poisoning or overdose	95 (51.1)	617 (74.0)
Self-injury (cutting, stabbing, self-battery)	74 (39.8)	171 (20.5)
Both self-poisoning and self-injury	3 (1.6)	29 (3.5)
Other - e.g. hanging, jumping from a height, running in front of transport	14 (7.5)	17 (2.0)
ICD-10 Axis 1 (pre or post first self-harm)	No. and prevalence of disorders (%)*	
Substance Misuse Disorders (F10-19)	10 (5.4)	13 (1.5)
Depressive disorder (F32)	53 (28.5)	277 (33.2)
Psychotic Disorders (F20-29,31,32.3, F33.3)	6 (3.3)	9 (1.1)
Anxiety Disorder (F40–42, F43–F48)	42 (22.5)	186 (22.3)
Eating Disorder	⩽5 (⩽2.5)	17 (2.0)
Autism Spectrum Disorders (F84)	18 (9.7)	21 (2.5)
Hyperkinetic disorder (F90)	19 (10.2)	15 (1.7)
Child-onset emotional and behavioural disorders(F91-F98)	33 (17.7)	127 (15.2)
No Diagnosis	41 (22.0)	249 (29.9)
Other	⩽5 (⩽2.5)	21 (2.5)
Axis 3 Intellectual Disability	⩽5 (⩽2.5)	6 (0.7)

*Multiple morbidities were counted, %

8.4.1 Characteristics of self-harm presentation

During follow-up, 1020 adolescents attended ED or were admitted to hospital with at least one episode of self-harm. Of these, 764 adolescents (~76%) were successfully matched to the National Pupil Database. Mean age of presentation was 15.9 years and 15.6 years for boys and girls respectively (see table 8.2). At the time of self-harm presentation, fewer than 50% had prior history of contact with CAMHS services. The most common reason for presentation was self-poisoning and overdose (50% of boys, 74% of girls), followed by cutting and self-battery. As shown in table 8.2, both boys and girls shared the same order and similar proportions of ICD-10 disorder prevalence, the most common being depressive disorders (Boy 29%, girl 33%) followed by anxiety (22% v 22%) and childhood onset emotional and behavioural disorders (such as oppositional and conduct disorders, 18% v 15% respectively).

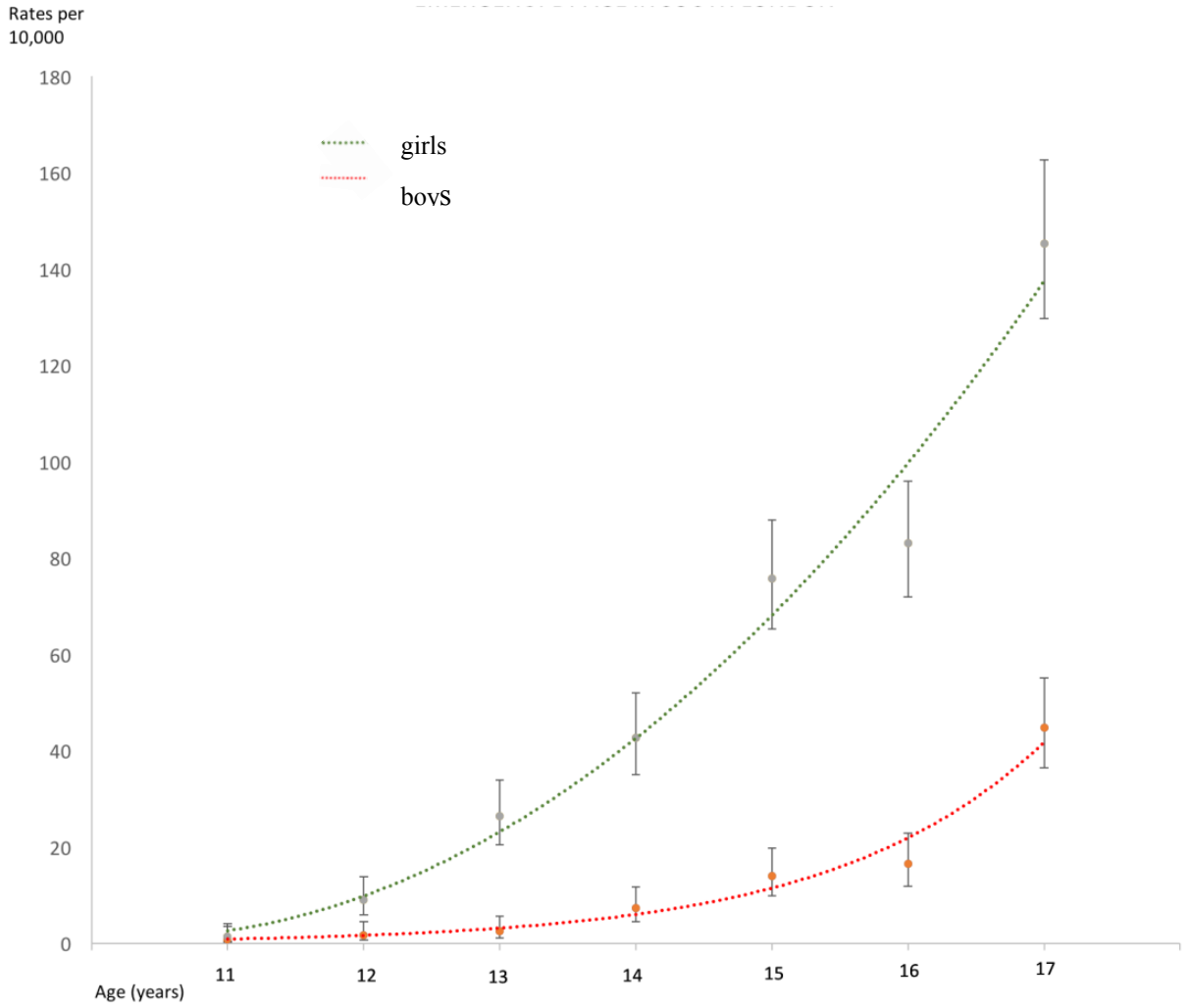
8.4.2 Incidence of self-harm by age and gender

Using individual-level NPD data to provide the regional population denominator, I assessed the incidence estimates of self-harm by gender. As described within the table nested in figure 8.3, and illustrated graphically. Both genders show low rates at age 11, with a substantial increase in incidence of self-harm throughout later adolescence. Incidence rates for 14 year girls were 42 per 10,000 increasing nearly four-fold to 145 per 10,000 at age 17. Although less in terms of absolute numbers, there were a greater relative increase for boys over this age range with rates at 14 years ~ 7 per 10,000 increasing six-fold to 45 per 10,000 at age 17.

8.4.3 Socio-demographic and education characteristics by gender and ASD status

Table 8.3 and 8.4 provide a breakdown of socio-demographic, educational and clinical characteristics of the sample, by gender and ASD status, provided by NPD (this data omits non-matched self-harm data, n=257). There was considerable ethnic, socio-economic and cultural diversity within the sample, with non-white ethnic groups making up over two-thirds of the study population, and over 25% reporting English as their second language. The majority resided in neighbourhoods within the highest 40% for national deprivation, with over 25% of adolescents coming from families meeting eligibility criteria for benefits or other income support.[340]

Figure 8.3 Self-harm incidence rates of adolescents presenting to A&E according to age and gender, with 95% CIs



Age (years)	Male			Female		
	n/pop at risk	per 10,000	95% C.I.	n/pop at risk	per 10,000	95% C.I.
11	2/23219	0.86	(0.2-3.4)	3/22979	1.31	(0.4-4.0)
12	4/23576	1.70	(0.6-4.5)	21/23229	9.04	(5.9-13.9)
13	6/23597	2.54	(1.1-5.7)	61/23168	26.33	(20.5-33.8)
14	17/23249	7.31	(4.5-11.8)	98/22995	42.62	(35.0-51.9)
15	32/22854	14.00	(9.9-19.8)	172/22714	75.72	(65.2-87.9)
16	36/21857	16.47	(11.9-22.8)	183/22038	83.04	(71.9-95.9)
17	89/19862	44.81	(36.4-55.1)	296/20372	145.30	(129.8-162.7)

Table 8.3 Socio-demographic, characteristics of the sample, by gender and ASD status

Socio-demographic characteristics**	Male (n=56,578)		Female (n=56,708)	
	No ASD (n=54,552)	ASD (n=2026)	No ASD (n=56,271)	ASD (n=437)
Mean age at baseline (SD)	12.8 (2.0)	12.3 (1.7)	12.8 (2.0)	12.2 (1.6)
Mean duration of follow-up (SD)	2.74 (1.3)	2.73 (1.3)	2.72 (1.3)	2.80 (1.3)
Ethnicity	(n, %)	(n, %)	(n, %)	(n, %)
White	20,238 (37.1)	770 (38.0)	20,651 (36.7)	176 (40.3)
Black	20, 012 (36.7)	850 (42.0)	21,099 (37.5)	174 (39.8)
Asian	4,788 (8.8)	78 (3.9)	4,869 (8.7)	24 (5.5)
Mixed	6,013 (11.0)	237 (11.7)	6,275 (11.2)	45 (10.3)
Other	1,928 (3.5)	42 (2.1)	1,891 (3.4)	6 (1.4)
not disclosed / unknown	1,575 (2.9)	49 (2.4)	1,486 (2.6)	12 (2.8)
National neighbourhood deprivation				
Most deprived quintile	19,805 (36.3)	824 (40.8)	20,222 (40.0)	173 (39.6)
2nd	22, 100 (40.5)	804 (39.8)	22,794 (40.5)	179 (40.1)
3rd	7,759 (14.2)	251 (12.4)	8,227 (14.6)	550 (12.6)
4th	3,283 (6.0)	99 (4.9)	3,322 (5.9)	24 (5.5)
Least deprived quintile	1,579 (2.9)	42 (2.1)	1,688 (3.0)	6 (1.4)

** Missing = 257 non-matched self-harm cases

There were 2,463 adolescents with an ASD registered special educational need, representing 2.2% of the total population. The majority of adolescents with ASD were being taught within mainstream schools (>75%), but 59-70% were in the lowest 20% for key stage 2 educational attainment (table 8.3), with 11-15% recognised as having severe or profound learning difficulties. For the ASD group, mean age was 12 years at study entry, with similar length of follow up (mean 2.7 years). Around 12% of boys and 6% of girls with ASD had received at least one fixed term exclusion. Between 5-6% did not attend school for more than 80% of the available lessons in the preceeding year before study entry.

Approximately 7% of boys and 5% girls with ASD had co-morbid hyperkinetic disorder, detected within the CAMHS record. 11 boys (0.5%) and less than 6 girls presented with self-harm at a mean age of 15 [statistical disclosure rules prevent actual numbers being provided].

Table 8.4 Educational and clinical characteristics of the sample, by gender and ASD status

Educational and Clinical Characteristics	Male (n=56,578)		Female (n=56,708)	
	No ASD (n=54,552)	ASD (n=2026)	No ASD (n=56,271)	ASD (n=437)
	(n, %)	(n, %)	(n, %)	(n, %)
Special Education Needs [a]				
Learning Difficulties (specific/moderate)	8,898 (16.3)	548 (27.1)	6,085 (10.8)	133 (30.4)
Learning Difficulties (severe/ profound)	591 (1.1)	250 (12.3)	381 (0.7)	65 (14.9)
Behavioural, Emotional, Social problems	6, 726 (12.3)	548 (27.1)	3,548 (6.3)	89 (20.4)
Speech, language and communication	4,291 (7.9)	806 (39.8)	2,134 (3.8)	161 (36.8)
Hearing, vision or physical disability	795 (1.5)	69 (3.4)	735 (1.3)	16 (3.4)
First language [a]				
English	39,920 (73.2)	1,661 (82.0)	40,815 (72.5)	344 (78.7)
Other	13,612 (25.0)	341 (16.8)	14, 541 (25.9)	89 (20.4)
Not disclosed	1,022 (1.9)	24 (1.2)	915 (1.6)	⩽5 (⩽1.0)
School Type				
Mainstream	53,868 (98.7)	1597 (78.8)	56,024 (99.6)	333 (76.2)
Special School	579 (1.1)	418 (20.6)	237 (0.4)	104 (23.8)
Pupil referral Units	107 (0.2)	11 (0.5)	10 (0.02)	⩽5 (⩽1.0)
Educational attainment (Key stage two) [b]				
Lowest quintile	12, 220 (23.1)	1,146 (59.0)	10, 461 (19.2)	296 (69.7)
second	10,461 (19.8)	277 (14.3)	10,750 (19.7)	55 (12.9)
third	10,301(19.5)	224 (11.5)	11,141 (20.4)	29 (6.8)
fourth	10,283 (19.5)	168 (8.6)	10,078 (20.3)	22 (5.2)
highest quintile	9,577 (18.1)	128 (6.6)	11,172 (20.4)	23 (5.4)
Less than 80% attendance [c]	2,587 (4.9)	118 (6.0)	2,538 (4.7)	22 (5.2)
Fixed term exclusions [a]	5, 847 (10.7)	239 (11.8)	2,790 (5.0)	26 (6.0)
Other social factors				
Summer birth (May -Aug)	18,941 (34.7)	720 (35.5)	19,185 (34.1)	140 (32.0)
Free school meals [a]	13,105 (24.0)	696 (34.5)	13,391 (23.8)	167 (38.2)
Looked after Child status [d]	420 (0.8)	30 (1.5)	397 (0.7)	12 (2.8)
ICD-10 Hyperkinetic disorder	670 (1.2)	131 (6.5)	168 (0.3)	23 (5.3)

missing values [a] 257 [b] 3731 [c] 4049 [d] 4547

Table 8.5 An analysis of socio-demographic risks factors for emergency presentations with self-harm amongst 113, 543 adolescents residing in south London using crude and multivariable cox-regression analyses.

Socio-demographic characteristics	Male (n=56,648)				Female (n=56,897)			
	No self-harm (n=56,462)	Self-harm (n=186)	Unadjusted Hazard Ratio	Adjusted Hazard Ratio	No self-harm (n=56,063)	Self-harm (n=834)	Unadjusted Hazard Ratio	Adjusted Hazard Ratio
Mean age at baseline (SD)	12.8 (2.1)	14.1 (1.8)	1.70 (1.55-1.86)**	1.38 (1.22-1.57)**	12.8 (2.0)	13.9 (1.8)	1.48 (1.42-1.54)**	1.28(1.21-1.35)**
Mean duration of follow-up (SD)	2.73 (1.3)	1.89 (1.2)	-	-	2.70 (1.3)	1.86 (1.1)	-	-
Ethnicity	(n, %)	(n, %)			(n, %)	(n, %)		
White	20,943 (37.1)	88 (47.3)	*reference*	*reference*	20,534 (36.6)	357(42.8)	*reference*	*reference*
Black	20, 842 (36.9)	28 (15.0)	0.32 (0.21-0.48)**	0.38 (0.23-0.65)**	21,106 (37.7)	212 (25.4)	0.57(0.48-0.68)**	0.58 (0.78-0.71)**
Asian	4,860 (8.6)	12 (6.5)	0.60 (0.33-1.10)	0.87 (0.35-2.14)	4,865 (8.7)	47 (5.7)	0.58 (0.43-0.78)**	0.61 (0.40-0.94)*
Mixed	6,234 (11.0)	176 (8.6)	0.62 (0.36-1.04)	0.69 (0.37-1.26)	6,218 (11.1)	102 (12.2)	0.97 (0.77-1.20)	1.12 (0.88-1.41)
Other	1,968 (3.5)	12 (6.5)	1.42 (0.78-2.60)	0.64 (0.14-2.69)	1,880 (3.3)	29 (3.5)	0.88 (0.60-1.28)	0.78 (0.46-1.31)
not disclosed	1,615 (2.9)	30 (16.1)	4.9 (3.3-7.5)**	0.74 (0.17-3.04)	1,460 (2.6)	87 (10.4)	4.0 (3.16-5.04)**	0.94 (0.54-1.61)
National neighbourhood deprivation [a]								
Most deprived quintile	20,586 (36.5)	63 (33.9)	*reference*	*reference*	20,144 (35.9)	320 (38.4)	*reference*	*reference*
2nd	22, 855 (40.5)	78 (42.5)	1.14 (0.82-1.58)	0.98 (0.63-1.53)	22,720 (40.6)	330 (39.5)	0.92 (0.79-1.08)	0.98 (0.81-1.17)
3rd	7,989 (14.2)	32 (17.2)	1.31 (0.86-2.20)	1.40 (0.81-2.42)	8,193 (14.6)	120 (14.4)	0.95 (0.77-1.17)	0.88 (0.67-1.15)
4th	3,378 (6.0)	8 (4.4)	0.78 (0.38-1.64)	0.74 (0.28-1.88)	3,311 (5.9)	44 (5.3)	0.85 (0.62-1.16)	0.80 (0.55-1.18)
Least deprived	1,620 (2.9)	5 (2.2)	0.81 (0.29-2.21)	0.27 (0.04-2.01)	1,677 (3.0)	20 (2.4)	0.75 (0.48-1.18)	0.79 (0.46-1.3)

* P⩽0.05 **P⩽0.01; [a] missing values= 52; [b] Adjusted for all other factors listed in this table and table 8.5

Table 8.6 An analysis of educational and clinical risks factors for emergency presentations with self-harm amongst adolescents residing in south London using crude and multivariable cox-regression analyses

Educational and Clinical characteristics	Male (n=56,581)				Female (n=56,709)			
	No self-harm (n=56,460) (n, %)	Self-harm (n=120) (n, %)	Unadjusted Hazard Ratio	Adjusted Hazard Ratio	No self-harm (n=56,063) (n, %)	Self-harm (n=646) (n, %)	Unadjusted Hazard Ratio	Adjusted Hazard Ratio
Special Education Needs [a]								
Autism Spectrum Disorders	2,015 (3.5)	11 (9.2)	2.73 (1.47-5.09)**	2.79 (1.40-5.57)**	434 (0.8)	⩽5 (⩽1.0)	0.57 (0.18-1.78)	0.52 (0.16-1.63)
Learning Difficulties (specific/moderate)	9,418 (16.7)	28 (23.3)	1.44 (0.95-2.20)	1.07 (0.62-1.76)	6,113 (10.9)	105 (16.3)	1.50 (1.22-1.85)**	0.99 (0.77-.127)
Learning Difficulties (severe/profound)	840 (1.5)	⩽5 (⩽5.0)	0.55 (0.08-3.92)	0.39 (0.05-2.98)	444 (0.8)	⩽5 (⩽1.0)	0.38 (0.09-1.52)	0.40 (0.10-1.67)
Behavioural, Emotional, Social	7,235 (12.8)	39 (33.5)	3.14 (2.19-4.70)**	1.66 (1.02-2.73)*	3,494 (6.2)	143 (22.1)	4.20 (3.48-5.05)**	2.31 (1.84-2.88)**
Speech, language and communication	5,086 (9.0)	11 (9.2)	1.06 (0.57-1.98)	0.99 (0.51-1.95)	2,269 (4.1)	26 (4.0)	1.01 (0.68-1.50)	1.13 (0.74-1.72)
Hearing, vision or physical disability	860 (1.5)	⩽5 (⩽5.0)	2.17 (0.80-5.89)	2.13 (0.77-5.85)	746 (1.3)	5 (0.8)	0.56 (0.23-1.34)	0.59 (0.25-1.42)
First language [a]								
English	41,482 (73.5)	100 (83.3)	*reference*	*reference*	40,652 (72.5)	508 (78.6)	*reference*	*reference*
Other	13,942 (24.7)	11 (9.2)	0.33 (0.18-0.62)**	0.50 (0.25-0.98)*	14, 529 (25.9)	101 (15.6)	0.57 (0.46-0.70)**	0.77 (0.61-0.98)*
Not disclosed	1,038 (1.8)	9 (7.5)	4.14 (2.10-8.2)**	n/a	882 (1.6)	37 (5.7)	3.82 (2.74-5.35)**	1.72 (0.91-3.02)
Educational attainment (Key stage two) [b]								
Lowest quintile	13,328 (24.4)	39 (33.0)	*reference*	*reference*	10,586 (19.5)	172 (27.0)	*reference*	*reference*
second	10,713 (19.6)	25 (21.2)	0.80 (0.40-1.32)	1.07 (0.60-1.90)	10,672 (19.6)	133 (20.9)	0.78 (0.62-0.97)*	1.01 (0.78-1.29)
third	10,501 (19.2)	24 (20.3)	0.82 (0.49-1.36)	1.56 (0.87-2.78)	11,046 (20.3)	124 (19.4)	0.73 (0.58-0.92)**	1.18 (0.90-1.52)
fourth	10,437 (19.1)	14 (11.9)	0.50 (0.27-0.92)*	1.01 (0.50-2.09)	10,974 (20.2)	126 (19.7)	0.77 (0.61-0.97)*	1.35 (1.04-1.77)*
highest quintile	9,9689 (17.7)	16 (13.6)	0.73 (0.41-1.31)	1.75 (0.85-3.55)	11,112 (20.4)	83 (13.0)	0.55 (0.44-0.75)**	1.15 (0.85-1.57)
Less than 80% attendance [c]	2,676 (4.9)	29 (26.4)	6.50 (4.24-9.92)**	3.50 (2.16-5.70)**	2,430 (4.5)	130 (21.2)	5.42 (4.50-6.58)**	2.84 (2.70-3.51)**
Fixed term exclusions [a]	6,054 (10.7)	32 (26.7)	2.88 (1.92-4.31)**	1.30 (0.78-2.15)	2696 (4.8)	120 (18.6)	4.41 (3.61-5.37)**	1.69 (1.32-2.15)**
Other social factors								
Summer birth (May-Aug)	19,615 (34.7)	47 (41.6)	1.21 (0.84-1.75)	1.23 (0.83-1.83)	19,104 (34.1)	222 (34.4)	1.02 (0.87-1.20)	1.02 (0.86-1.21)
Free school meals [a]	13, 764 (24.4)	37 (30.8)	1.40 (0.95-2.05)	1.35 (0.87-2.10)	13,369 (22.1)	189 (29.3)	1.32 (1.11-1.56)**	1.22 (1.02-1.48)*
Looked after Child status [d]	443 (0.8)	7 (6.3)	8.04 (3.75-17.3)**	3.18 (1.14-8.91)*	382 (0.7)	27 (4.3)	6.20 (4.22-9.12)**	3.16 (2.07-4.84)**
ICD-10 Hyperkinetic disorder	788 (1.4)	19 (10.2)	8.0 (5.0-12.8)**	4.36 (2.20-8.68)**	177 (0.3)	15 (1.8)	5.70 (3.42-9.50)	3.58 (2.03-6.29)**

* P⩽0.05 **P⩽0.01; missing values [a] 257 [b] 3731 [c] 4049 [d] 4547 [e] Adjusted for all other factors listed in this table and table 8.5

8.4.4 Population level socio-demographic and educational risks for self-harm by gender

Cox regression models displayed in table 8.5 indicate, for both boys and girls, a strong inverse association between black ethnicity (relative to white ethnicity) and risk of presenting with self-harm. This association remained significant, and the effect estimate consistent, even after adjustment for a range of potential confounders, including clinical and educational factors. Asian ethnicity and English as a second language (i.e. English not the primary language spoken at home), were also associated with significantly reduced risks of self-harm presentation, but only amongst girls. Levels of neighbourhood deprivation were not significantly associated with risk of self-harm for either gender.

Table 8.6 shows the educational and clinical risk factors associated with self-harm, stratified by gender. ASD was associated with nearly a three-fold increase in risk self-harm in boys, showing little change after adjustment for a comprehensive range of clinical confounders (aH.R 2.79, $P<0.01$), however ASD was not an associated risk for girls ASD. Other significant predictors for self-harm in both genders included behavioural, emotional and social special educational needs, persistent attendance problems, being a looked after child, and hyperkinetic disorder. For girls specifically, being from a family eligible for free school meals, having at least one fixed term exclusion from school, and being in the second from top highest achieving academic quintile were also significant predictive factors.

8.4.5 Sensitivity analysis

Previously specified sensitivity analyses made little difference to the main findings. Restricting the analyses to adolescents joining the study aged 11, showed that ASD in boys remained a significant risk factor (aH.R 3.43, 95% C.I. 1.05-11.3, $p<0.04$), restricting to those enrolled in mainstream school produced similar results (aH.R 3.28, 95% C.I. 1.64-6.6, $p<0.01$). The final analyses used an imputed dataset, which replaced missing NPD variables that were either not supplied to Department for Education, or missed matches between NPD and CRIS data. Table 8.7, shows the distribution of key variables before and after multiple imputation, which I checked to establish the validity of this imputed dataset. Observed values of complete cases with imputed values showed similar distributions, with the exception of a nearly 3-fold increase in the proportion of adolescent who did not disclose their language status. Table 8.8 shows fully adjusted effect estimates are similar to the complete case analyses in table 8.5 and 8.6, except with some gains in precision.

Table 8.7 The distribution of socio-demographic and educational variables before (original) and after multiple imputation.

Socio-demographic and clinical characteristics	Male (n=56,648)				Female (n=56, 709)			
	Original %		Imputed %		Original %		Imputed %	
	No self-harm	Self-harm	No self-harm	Self-harm	No self-harm	Self-harm	No self-harm	Self-harm
National neighbourhood deprivation								
Most deprived quintile	36.5	33.9	33.6	37.5	35.9	38.4	32.8	37.5
2nd	40.5	42.5	40.8	38.1	40.8	39.5	40.6	38.1
3rd	14.2	17.2	15.4	16.3	14.6	14.4	16.3	17.2
4th	6.0	4.4	7.0	7.2	5.9	5.3	7.2	5.4
Least deprived quintile	2.9	2.2	3.2	3.3	3.0	2.4	3.3	2.2
Special Education Needs [a]								
Autism Spectrum Disorders	3.5	9.2	3.5	5.1	0.8	0.5	0.7	0.8
Learning Difficulties (specific/moderate)	16.7	23.3	14.4	24.4	10.9	16.3	9.3	18.5
Learning Difficulties (severe/profound)	1.5	0.8	1.4	1.0	0.8	0.3	0.7	0.4
Behavioural, Emotional, Social problems	12.8	33.5	12.5	41.4	6.2	22.1	6.0	23.2
Speech, language and communication	9.0	9.2	7.5	8.3	4.1	4.0	3.4	3.9
Hearing, vision or physical disability	1.5	3.3	1.3	1.6	1.3	0.8	1.3	0.8
First language [a]								
English	73.5	83.3	72.0	66.5	72.5	78.6	72.1	68.8
Other	24.7	9.2	23.0	11.3	25.9	15.6	25.6	13.8
Not disclosed	1.8	7.5	5.1	24.2	1.6	5.7	4.3	17.5
Educational attainment (Key stage two) [b]								
Lowest quintile	24.4	33.0	23.2	39.6	19.5	27.0	17.7	29.6
second	19.6	21.2	18.0	20.0	19.6	20.9	18.3	21.6
third	19.2	20.3	18.5	14.3	20.3	19.4	19.1	17.7
fourth	19.1	11.9	19.9	13.7	20.2	19.7	21.3	16.9
highest quintile	17.7	13.6	20.5	12.4	20.4	13.0	23.6	14.3
Less than 80% attendance [c]	4.9	26.4	5.5	31.3	4.5	21.2	5.0	25.7
Fixed term exclusions [a]	10.7	26.7	10.9	37.8	4.8	18.6	5.3	23.2
Other social factors								
Summer birth (May -Aug)	34.4	41.6	34.7	36.3	34.1	34.4	33.8	34.5
Free school meals [a]	24.4	30.8	21.9	32.5	22.1	29.3	20.1	30.2

Table 8.8 An analysis of educational and clinical risks factors for emergency presentations with self-harm using multiple imputed data.

Socio-demographic, educational and clinical characteristics	Imputed Sample	
	Male	Female
	Adjusted Hazard Ratio	Adjusted Hazard Ratio
Mean age at baseline (SD)	1.63 (1.47-1.80)**	1.36 (1.31-1.42)**
Ethnicity		
White	*reference*	*reference*
Black	0.38 (0.24-0.58)**	0.60 (0.50-0.72)**
Asian	1.23 (0.63-2.39)	0.87 (0.63-1.20)
Mixed	0.61 (0.35-1.04)	0.95 (0.76-1.18)
Other	2.60 (1.34-5.01)**	1.18 (0.79-1.77)
not disclosed / unknown	2.89 (0.17-3.04)**	2.21 (1.57-3.10)
National neighbourhood deprivation [a]		
Most deprived quintile	*reference*	*reference*
2nd	1.17(0.85-1.62)	0.96 (0.82-1.12)
3rd	1.32 (0.85-2.06)	0.98 (0.79-1.23)
4th	0.76 (0.35-1.62)	0.83 (0.60-1.15)
Least deprived quintile	0.76 (0.27-2.07)	0.76 (0.48-1.21)
Special Education Needs [a]		
Autism Spectrum Disorders	2.32 (1.28-4.26**)	0.62 (0.16-1.63)
Learning Difficulties (specific/moderate)	1.16 (0.83-1.89)	1.11 (0.88-1.39)
Learning Difficulties (severe/profound)	0.41 (0.05-3.14)	0.44 (0.11-1.77)
Behavioural, Emotional, Social problems	2.02 (1.27-3.22)**	2.08 (1.67-2.58)**
Speech, language and communication	1.07 (0.53-2.15)	1.16 (0.77-1.74)
Hearing, vision or physical disability	1.58 (0.60-4.15)	0.57 (0.23-1.41)
First language		
English	*reference*	*reference*
Other	0.48 (0.25-0.85)**	0.70 (0.55-0.89)**
Not disclosed	0.98 (0.33-2.86)	1.37 (0.93-2.05)
Educational attainment (Key stage two)		
Lowest quintile	*reference*	*reference*
second	1.09 (0.70-1.69)	1.05 (0.80-1.37)
third	1.33 (0.76-2.43)	1.18 (0.93-1.51)
fourth	1.11 (0.58-2.13)	1.32 (0.99-1.74)
highest quintile	1.76 (0.96-3.23)	1.17 (0.87-1.57)
Less than 80% attendance [c]	3.03 (1.87-5.01)**	2.67 (2.18 -3.26)**
Fixed term exclusions	1.25 (0.83-1.89)	1.60 (1.26-2.01)**
Other social factors		
Summer birth (May -Aug)	1.20 (0.83-1.72)	1.01 (0.86-1.19)
Free school meals [a]	1.30 (0.85-1.97)	1.27 (1.07-1.50)**
Looked after Child status [d]	3.71 (1.88-7.35)**	2.78 (1.94-3.97)**
ICD-10 Hyperkinetic disorder	3.96 (2.27-6.93)**	3.22 (1.91-5.43)**

*$P<0.05$, **$P<0.01$

8.5 DISCUSSION

To my knowledge, this is the first investigation to examine whether ASD is a population level risk factor for emergency department presentation of self-harm and in adolescence. Using electronic health record data from a comprehensive specialist child mental health care service, supplemented by accident and emergency attendance records, and linked to longitudinal school records of all pupils within a defined geographic catchment, I found evidence that ASD was associated with nearly 3-fold increased risk of self-harm among boys. This association persisted after controlling for a broad range of potential confounders[246] including socio-economic and demographic and factors, learning difficulties, academic attainment, educational markers of emotional and behavioural severity, school exclusion, reduced attendance, childhood maltreatment, hyperkinetic disorders. It was also robust to sensitively analyses to reduce the heterogeneity, and potential residual confounding factors within the cohort.

Our findings are consistent with the limited research conducted that show adolescents with ASD are at greater risk for reporting suicidal behaviours.[150,151] However, these studies have not provided evidence that ASD is a population level risk factor for suicidal behaviours. They were limited by reliance on clinical populations with ASD sampled from mental health services, who were more likely to present with greater levels of psychiatric need than those diagnosed and managed within community paediatric and specialist educational settings.[341] I found ASD was a significant risk factor for adolescent boys only, however this should not be taken to imply that boys with ASD are at greater risk than girls with ASD. The rate of self-harm amongst girls with ASD were similar to boys (statistical disclosure rules do not permit publication of the actual figures), and as shown in figure 8.3 self-harm incidence rates were far higher amongst girls.

The gender discrepancy found in this study may be seen as inconsistent with recent findings from Hirvikoski et al, who found adult women with ASD were 13 times more likely to die from suicide, compared to the 6-fold risk found in males.[291] There may be several potential explanations for this. While the natural course of self-harm shows that population rates decrease significantly in girls over early adulthood, rates of self-harm in adolescent ASD girls may not naturally decline, and go on to contribute to later risk of adult suicide. ASD is an under-recognised condition among girls within paediatric and school settings, so another possibility is that, detection/diagnosis of ASD among adult women will largely be through self-presentation to psychiatric services.[291,322] Adult ASD presentations in this context will be strongly associated

with distress and psychiatric co-morbidity, particularly severe anxiety and depressive disorders,[322] which in turn are strong predictors for suicide. One additional explanation is that a number of girls who self-harmed within this study also had undiagnosed ASD, and were included within general population rates, producing an underestimate of the true effect of ASD on self-harm due to misclassification.

Self-harm presentation could be perceived as a potential proxy for failing to respond and address adolescent psychopathology sufficiently. Particular issues with addressing this need among people with ASD may explain the elevated risk that I have identified. Emerging evidence shows risk factors and treatment targets for psychopathology established in neuro-typical populations may not translate onto those with ASD. For example, prevalence rates and causes of common non-neurodevelopmental psychiatric disorders, such as oppositional-defiance, depression or anxiety from non-ASD populations, differ when examined as co-morbidities within ASD.[118,299] Furthermore, ASD related characteristics such as difficulties with social reciprocity, social communication, flexibility, sensory processing and emotional recognition may only precipitate maladaptive responses and psychiatric co-morbidity when new environmental challenges arise.[342] Ecological shifts such as moving up to secondary school,[343] or physiological changes such as entering puberty [344] may specifically interact with these social difficulties, overwhelming the functional capabilities of adolescents with ASD and causing significant psychiatric impairment.

It cannot be assumed that conventional approaches to detecting mental health problems are effective in ASD. Social communication is almost always impaired in ASD, making it difficult for family members, clinicians or the individual's themselves to recognise changes in their emotional states and seek help. To add to this complexity, psychiatric symptoms such as anxiety, self-injury, hyperactivity or disruptive behaviours were once viewed as part of the social and behavioural characteristics of ASD.[118,345–347] These factors have contributed to a legacy of under recognition and unmet psychiatric need among people with ASD, and so psychiatric assessments and therapeutic tools tailored to those who have ASD are novel and lack robust evaluation. If mental health needs are unmet for children and adolescents with the ASD, as with the general population, they may have broad, and enduring implications for immediate and later quality of life. [348,349] Recent priority setting exercises for ASD recognise this, and have designated as the highest priority research which aims to delineate which mental health problems arise, and how they are best treated.[350] Given difficulties in identification, higher rates of psychopathology and

a limited understanding of the treatment targets for treating psychiatric co-morbidity, it could be expected that adverse consequences of psychiatric co-morbidity such as self-harm are greater in ASD populations relative to the general population.

Other findings deserve comment. I found a robust longitudinal association in both boys and girls, showing persistent absence from school at baseline (in the year prior to study entry) was associated approximately with a 3-fold increase in self-harm. As far as I am aware this is first population-based longitudinal study describing such an effect. These findings alone do not show absenteeism directly causes mental health difficulties - I was unable to ascertain the reasons for persistent absence, such as truancy, school refusal, or health problems not captured by special education need categories. That said, the findings certainly show persistent absence is a strong signal for vulnerability and later psychopathological disturbance, and, I believe, provides robust evidence to support routine screening for potential unmet mental health needs in young people with low attendance rates. Study findings of school exclusion, and behavioural, emotional and social special educational needs (BESN) predicting later self-harm were consistent with a small scale cross-sectional study showing significantly higher rates (22%) of self-harm amongst adolescents with a history of exclusion or BESN.[351] In addition, I found ADHD, a condition not specified within any SEN category, was a strong predictor for self-harm. The study showed ADHD is associated with approximately a four-fold risk for self-harm for both genders, and hence addressed a gap in the evidence base with very few prospective studies addressing the psychiatric consequences of ADHD, particularly in girls. [352]

In the adjusted model, I found free-school meal eligibility (a proxy for low socio-economic status) was significantly associated with self-harm in girls. This finding was consistent with a number of studies examining social-economic factors in self-harm.[353,354] I found both boys and girls, compared to children ineligible for free school meals, were at 30% greater risk for self-harm, however the strength of the association for boys did not reach statistical significance, which may relate to lack of power given the very smaller number of boys who presented with self- harm. I found an interesting effect of primary school academic attainment on adolescent self-harm. In the non-adjusted analysis, educational attainment was inversely correlated with self-harm, however after comprehensive adjustment for a number of factors, including special education needs and a range of behavioural factors, I found being in the second highest attainment level was positively correlated with self-harm. This could be a chance finding or

secondary to residual confounding possibly through failing to take account of internalising disorders, which are associated with self-harm in adolescence.[355]

I found peak incidence rates for both genders occurred at age 17, with boys at 44 per 10,000 and girls at 145 per 10,000 population. I found self-harm incident rates increased significantly by age, over the adolescent period for both boys and girls. These findings are consistent with a number of studies using hospital data, which also show girls have higher rates of self-harm than boys, over 11-17 year age range.[311,356,357] The findings suggest that self-poisoning alone was the most common form of self-harm in adolescence with 74% of girls and 51% of boys having this as an identified reason for presentation, followed by self-injury, girls (21%) and boys (40%). These results show a similar pattern and frequency to two large scale hospital based UK studies.[356,357] Consistent with previous studies,[246] I found very high rates of psychiatric morbidity in adolescents presenting with self-harm. Approximately 50% of adolescents who presented to ED had not received any specialist mental health support prior to self-harm presentation, again suggesting considerable unmet psychiatric need within this region of South London.

Consistent with other hospital based studies, the rates presented are likely to represent a fraction of self-harm within the adolescent community. Many well conducted surveys have reported the annual prevalence of adolescent self-harm at around 8-12%.[316,329,358] However, using serial school census data as a population as denominator for the region, combined with free text extraction, and case note review of the mental health record at the time self-harm presentation, I found far higher rates than the published figures derived from self-harm inpatient admission rates within national surveillance figures in HES. Annual self-harm rates between 2014-2015 for the region covered in this study, were 22.8 per 10,000 for adolescents aged between 10-24 years.[359] These findings are comparable to a recent French study, which also found free text extraction of self-harm from ED medical records produced significantly higher rates than public health surveillance systems.[95] This discrepancy found in my study is also consistent with prior work comparing routinely collected admission data on self-harm to research data for the same regions, where routine admission data was found to underestimate self-harm presenting to emergency departments by up 60%.[245]

Rates of admission related to self-harm are used as the indicator to represent mental health and well-being in Public Health England's National Child and Maternal Health Intelligence Profiles.[360] However, these statistics have limitations. They can only represent the proportion of

self-harm that results in a hospital admission, and so miss presentations with self-harm that are seen and discharged from emergency departments without requiring admission. In the UK, National Institute for Health and Care Excellence guidelines currently recommend that a young person (aged under 18) who self-harms is admitted to a paediatric ward and that a professional skilled in assessing mental health problems in young people carries out a psychosocial assessment within 24 hours.[332] However, from the limited evidence available, in clinical practice only 30% of older adolescents (aged 16+) and young adults (aged 18-25) are admitted to general hospital which is no more likely than older adults presenting with self-harm.[361] Another UK based study found younger adolescents (aged 12-14) were more likely to be admitted to a paediatric unit, but again, even at this age, a general hospital inpatient stay was not inevitable.[357] This suggests that a significant proportion of adolescent self-harm presentations receiving emergency medical attention are not included in HES inpatient data.

8.5.1 Strengths

This study has a number of strengths. Using routinely collected data from schools, I was able to ascertain longitudinal follow-up data on a very large population based sample, with participation and retention of many individuals at risk who traditionally may be lost to follow-up.[44] Samples were sufficient size and adequate statistical power to conduct robust analyses, allowing adjustment for a range of potential confounders which addressed the sample size and/or measurement limitations of previous studies. The main finding was robust to a number of sensitivity analyses, including potential differential selection caused by data linkage errors. The data linkage and extraction strategies within clinical notes enabled self-harm outcomes to be collected as objective endpoints and hence less subject to information biases recall and observer bias. The NLP approaches provided detailed clinical information held with free-text notes, permitting validation work to be conducted, to ensure that the appropriate self-harm construct was ascertained. These data would not normally be available in an electronic case register derived purely from structured, administrative healthcare data. Also, the linked data with education, improved on the conventional health database studies of self-harm, which, in terms of detecting school based risk factors, have been limited in their scope as they not been able to capture school data, and therefore not included these key risks in analysis models. Finally, the study has provided a novel, but replicable methodology to other UK regions, and gives an example of how large scale epidemiological approaches to examining self-harm patterns and risks in adolescence can be enhanced.

8.5.2 Limitations

The findings need to be interpreted in the light of several limitations. First, is the lack of pupil level measures for internalising problems, which are known to co-occur with externalising disorders [362] to a greater extent in lower academic achievement groups.[363] This analysis took account of the effects of more severe externalising problems, for example through adjusting for ADHD, behavioural and conduct problems SEN categories, and the school exclusion variables. If there was differential collinearity between externalising and internalising disorder according to academic attainment, the model may have provided greater adjustment of internalising psychopathology in lower academic groups – removing externalising problems and collinear internalising effects on self-harm - leaving internalising disorders as residual factor in the higher achieving groups. This may drive the association I found between a higher attainment group and self-harm. This explanation is tentative, as evidence for the longitudinal associations between academic attainment and internalising symptoms remain inconclusive.[363,364] However, recent work by Patalay et al, showed in a UK primary school sample, the group with worsening internalising trajectory developed the greatest level of psychopathology by end of follow up, also had the highest mean scores in primary school education attainment. [365]

I was unable to completely capture exposure variables for the whole population at risk, and it is unlikely, given my prior work on data linkage studies, that data were not missing completely at random. This non-random response may have biased the complete case analyses. However, the complete case analyses and subsequent analyses using imputed data were consistent, suggesting that biases, which may have arisen in the complete case analyses, did not significantly affect the study findings.

I only had 4 years of outcome data available for four boroughs, and because of the way I captured the baseline population via the NPD, I could not extend the analysis to include young adults, who are another high-risk population but do not have their area of residence collected annually by the educational database. The restricted catchment area meant that adolescent residents who presented to hospitals outside the catchment, and were not admitted, may have been missed. The restricted time period also meant I had insufficient power to examine self-harm for girls with ASD, nor examine those who later progressed to severe suicide attempts. Furthermore, I was not able to distinguish those who had self-harmed with suicidal intent from those who had self-

harmed without suicidal intent. However, building on my previous work using natural language methodologies (see chapter 3),[366] I aim to capture suicidal intention within the free text records, alongside self-harm, and examine its combined predictive validity as a risk factor for suicide attempts. The study was limited to identifying risk factors for self-harm. Conversely, to better inform mental health interventions and preventive work, future work should also focus on which factors are associated with adolescents, with and without ASD, who stop self-harming, which I hope to address in subsequent follow-up of this cohort. As with all observational studies, there is also the possibility of residual confounding, whereby associations may be accounted for by an additional unmeasured variable.

8.5.3 Conclusion

This study was dependent on using linked de-anonymised free-text electronic health record, education and hospital administrative data collected from schools, mental health and acute hospital trusts. I believe this is a first for UK based child and adolescent mental health research, and an important example of how data linkage work can be used to tackle important public health issues. If a similar research programme was developed to evaluate adverse outcomes associated with ASD, using more conventional cohort designs, it would have taken up significant time and resources to deliver a representative population and sufficiently powered analysis. Furthermore, given the dynamic nature of public services changes, demographic shifts in sample populations [156], youth education [367] and mental health policies and practice, by the time data has been collected, analysed and published, the findings have the potential to become quickly outdated. [244] Because the linkage and detection techniques for the data resources used in the study can be completely automated, the methodology described in this study could be extended to other areas of the UK to improve contemporaneous self-harm detection rates from routinely collected data, and provide better intelligence for future resource allocation.[156] The continued debate and engagement with patients, researchers and the wide public on the use of routinely collected data held by public services helps support the research conducted in this chapter. I hope that this work, along with evidence from other groups using de-anonymised linked public service data, helps provide an example of how this data can be used to public benefit.

In summary, this study has identified ASD as a subgroup of adolescents who have a pronounced risk of self-harm, and provides an important step towards addressing which factors within ASD lead to high lethality suicide attempts, and premature mortality.

CHAPTER 9. DISCUSSION AND CONCLUSIONS

This thesis examined original and clinically relevant research questions using data from routinely collected clinical text, enriched by NLP and linkages to external data sources, and nested within a local population. The objective of this work was to examine how data linkage and NLP approaches could expand the comprehensiveness of information available in child and adolescent mental health records for analyses and hypothesis testing. Five related studies were performed (covered in chapters 2-8), all using data obtained from the SLaM BRC Clinical Records Interactive Search (CRIS) extracted using a NLP approaches, chapters 6-8 describing studies using external linkages with routinely collected national electronic datasets (HES and NPD). The programme of work set out in the thesis, similar to the technologies employed in the study methodologies, have developed iteratively. Each study has addressed some aspect of the methodological limitations identified within the study described in chapter 2. Individual discussions and conclusions were presented within each of the study chapters (2-8). The current chapter summarises the key findings from the studies and presents the overall strengths and limitations of the work, before outlining its potential implications and contributions and future directions for research.

9.1 SUMMARY OF THESIS

9.1.1 The impact of child and adolescent psychiatric co-morbidity on antipsychotic treatment and outcomes

Pharmaco-epidemiological studies have demonstrated that psychiatric medication use in child and adolescent clinical populations has been increasing exponentially (see chapter 2). Of particular concern is the growing trend of antipsychotic medication use in ASD, especially over its use as an 'off-label' treatment for psychiatric conditions in ASD. Prescribing practices vary considerably across the world, and most evidence is derived from the United States, which has different licensing arrangements and practices for antipsychotic prescribing compared to the UK. The investigations contained in this thesis provide novel contributions to the evidence base, by highlighting the clinical factors, especially co-morbid psychiatric disorders, that are associated with the use of antipsychotics in ASD (chapter 2). This work provides much needed pharmaco-surveillance on not just rates of prescribing, but also indications for their use. Prior to this work, no longitudinal studies had examined challenging behaviours and psychiatric comorbidity

profiles as predictors of antipsychotic use in ASD. Most studies were limited by relying on parental recall of past comorbidities and medication use, retrospective or cross-sectional design, or the confounding effects of unmeasured psychiatric symptoms and disorder severity not being accounted for.[106,107,142]

Using a different clinical population, the thesis also contributed to the literature by examining the predictive factors for antipsychotic treatment failure, or MTF, in adolescents with early onset psychosis (chapter 4 and 5). These were the first studies in adolescent samples I am aware of, which examined the association of ASD and NS phenotypes as predictors for either antipsychotic treatment failure, or any related refractory treatment outcome like treatment resistance.[368] The findings suggest that ASD and NS profiles could be phenotypic markers for adolescent psychotic disorders which are harder to treat with conventional antipsychotics, and therefore result in a more impaired illness course.

The work in this thesis initially suggested that some psychiatric conditions, within the context of being co-morbid with ASD, may also lower the risk of antipsychotic treatment compared to non-ASD children with the psychiatric condition (chapter 2). For example, I interpreted the finding that only 47% of children with ASD and co-morbid psychosis received antipsychotics, and suggested that this was a lower treatment rate than expected in non-ASD child population with psychotic disorder. I attributed this low rate to potential diagnostic uncertainty between ASD and psychotic symptoms (chapter 2), and that clinicians may decide that some psychotic symptoms within ASD do not warrant antipsychotic treatment. However, as the thesis evolved, a potentially more complex picture developed, where I found ASD was a risk factor for multiple antipsychotic failure (chapter 4). My initial assumption may have been potentially incorrect due a *floating numerator* error, where I did not have an unexposed comparison group. [369] This highlighted the importance of using appropriate control groups (as illustrated in chapter 4 and 5), and was also taken up further in chapters 6 and 8.

9.1.2 Enhancing observational study approaches in child and adolescent psychiatric epidemiology: using NLP tools in health records

This thesis demonstrated how NLP procedures can be used with EHRs to better extract risk factor and outcome data for analysis, and address important research questions in child and adolescent mental health. As highlighted (chapter 1), a key limitation of studies using purely administrative health record data are that the variables coding risk factors and outcome factors, are extracted

from structured fields, and lack further contextual information. These limitations were also highlighted in my work, when I just used the structured risk item for self-injurious behaviours (chapter 2), which provided no indication why a child was scored at high risk of injuring themselves.

I show in this thesis how the limitations of structured data extraction, can be addressed by using more intricate text extraction methodologies to better assess symptom types, severity and related impairments, for example with suicidal risk (chapter 3), negative symptoms (chapter 5) and self-harm (chapter 8). In the thesis, I developed and evaluated an NLP approach which could accurately identify suicidality (chapter 3) in young people with ASD, using the adolescent sample in chapter 2. This was the first study to demonstrate that a NLP tool can be used to accurately capture a clinical construct as complex as suicidality within health records of young people with ASD. The NLP tool identified suicidality-related mentions with high degrees of precision and recall from clinical free text held within EHRs, and demonstrated that NLP applications can provides powerful opportunities for surveillance work of suicidality in adolescent ASD and in other clinical samples. However, an issue with the NLP tool I developed was that it lacked the ability to address temporality: target terms related to suicidality were not contextualised in terms of historicity. So, the tool was able to describe the prevalence of suicidality in the ASD clinical sample but could not capture suicidality incidence, limiting its potential for providing clinical outcome data in longitudinal study designs. However, in the thesis I went on to demonstrate an approach for addressing the issues around temporality. Accurate event data within the structured fields, such as date of first presentation to services (chapters 4, 5 and 8), was used to define periods of time for NLP extraction of exposures (negative symptoms, chapter 5) and outcomes (self-harm, chapter 8).

9.1.3 Enhancing observational study approaches in child and adolescent psychiatric epidemiology: Combining multiple sources of public service data.

In this thesis, I made a clear case for local areas to link existing routinely collected data created by public services within their region, to enhance their ability to conduct population-based analyses on clinical outcomes (chapter 6). Similar to floating numerator error in epidemiology, caused by lack of an appropriate denominator, I argued that without reference to local area's population, it was difficult to assess how well services are meeting the local needs if only using clinical data.

For example, a clinical service may appear to provide exemplary care to all children referred with ASD, but without a population perspective of ASD across the region, the service may only be serving a sub-group or a small proportion of those eligible; so its overall impact for the ASD population may be very limited.

To provide a case example for how data linkage methods can inform local population-based analyses, I developed a linkage between NHS child mental health and education data, with the aim that the data acquired should be a valuable enhancement to child-based longitudinal studies and clinical registries (chapter 7, 8). The linked school and health data supplied appropriate denominators for population-based analyses at relatively low cost, allowing evaluation of questions relevant to public health and social care policy.

My research has shown that combining NLP approaches to exploit EHRs data linked to public service data is a powerful approach for examining outcomes of rare exposures like ASD and rare outcome events, like self-harm, which are insufficiently captured by current NHS systems (see chapter 8). In this thesis, I conducted a population-based study using Department of Education data to provide a whole region sample of individuals attending school and their sample characteristics linked to NHS mental health data. Using these data, I determined that for boys, ASD was a population level risk factor for presenting to hospital with self-harm. I also reported a number of novel education-based risk factors for self-harm, which included persistent school absence and school exclusion.

I demonstrated that the legal, governance and technical challenges are surmountable and described the first study in England to link NHS and Department for Education routinely collected school's data together (see chapter 7). There are a number of lessons to be learned about how legal frameworks in England are applied when seeking to link routinely collected health information to other public service data without individual consent. For example, when attempting to gain section 251 approval (see chapter 7), it may be challenging understanding the legal definitions of medical purpose, where the current interpretations of section 251 legislation definitions are not necessarily intuitive to child psychiatry.[264]

The work conducted demonstrated that a key justification for using a 'opt out' governance approach - not relying on individual level consent to link data - was to study groups that traditionally were hard to reach and retain in observational studies (see chapters 6 and 7).[44] The

results indicated, the opt-out consent approach used may have improved representation of more socially disadvantaged populations. I found no difference in linkage rates between those from the highest and lowest quartiles of neighbourhood deprivation (see chapter 7). Nevertheless, whether using opt in or opt out consent process, possible biases due to linkage error can be substantial and need to be examined when analysing and interpreting results. Differential linkage error by ethnicity, social disadvantage and clinical factors can introduce significant selection bias leading to inaccurate risk factor-outcome estimates, which in turn may have significant impact on the validity of the research findings using the linked data. However, the findings in chapter 7 demonstrated that exposure-outcome associations between health and education factors, may be robust to linkage biases even with incomplete linkage between datasets and differential linkage success across socio-demographic groups.

The thesis provided an example of how non-random loss between routinely collected health and non-health linked data can be adjusted by weighting techniques (chapter 7 and 8). It showed that this approach may be useful to determine whether the reported associations were effected by linkage error lead to systematic bias caused by the linkage techniques. My research indicated the importance of data sharing agreements and the relationship between linkers, data controllers and analysts being sufficiently developed to share information on linked and unlinked data from source files, in order to appropriately adjust for linkage error when it occurs (see chapter 7).

9.2 STRENGTHS

A strength of all the studies reported in this thesis was derived from the use of large scale child health data held within the SLaM electronic patient record. All the studies described in chapters 2 to 8, involved distinct data collections on clinical samples ranging between 1000 and 35,500 patients. NLP approaches enabled these large samples to be sufficiently characterised, and used to address research questions which would have otherwise been unfeasible to investigate in observational studies involving direct patient recruitment. The samples were of sufficient size and adequate statistical power to conduct robust analyses, allowing adjustment for a range of potential confounders which helped address the sample size and/or measurement limitations of previous studies.

Another advantage of using the child health record data in SLaM, is that the data captured the total clinical population of interest. As described in chapter 6, SLaM holds a monopoly over CAMHS provision for its geographic catchment, hence the clinical record captures patients accessing both inpatients and community settings. Studies which are able to combine sources of data from both setting are rare, especially in pharmaco-epidemiological studies.[55,368]

Crucially, the studies in this thesis have used electronic records to capture key clinical outcomes which were very likely to come to the attention of local specialist mental health services, or, in terms of educational outcomes, be captured systematically in educational administrative systems. In addition, the linkage of CRIS and NPD systems are based on an 'opt out' rather than opt-in governance model and to date only three patients (no caregivers or young people) have asked for their records to be removed from the CRIS search system. So, with near 100% coverage of all young people receiving specialist child mental services and no discernible consent bias, the analyses in thesis should be less susceptible to risk of selection and loss to follow-up biases. Especially those biases that are associated with conventional recruitment and measurement approaches used in surveys, cohort studies and randomized controlled trials. Primary data collection often used in these studies can result in reducing clinically representative samples, because those clinically relevant characteristics, especially in mental health studies, are also associated with low participation and a reduced chance of being selected into research studies. [370] That said, this is not always an advantage of large sample sizes and "big data" studies. Despite providing precise estimates on the effects of potential risk factors on clinical outcomes, these studies are of little value if the sample are not representative of the population, or missing key information on a non-random basis.[371] This is a strength of studies described in chapters 7 and 8, where adjustments to reduce the impact of linkage error were made either through statistical adjustment / weighting (chapter 7) or through imputing exposure variables via multiple imputation (chapter 8). These efforts were made to reduce the potential selection bias of the big data approaches, and support the representativeness of the study findings.

Another strength was that the data linkage and extraction strategies within clinical notes enabled outcomes to be collected as objective endpoints and hence less subject to information biases, including recall and observer biases. The outcomes included in this thesis were antipsychotic prescribing within child and adolescent samples with ASD, psychosis (chapters 2, and 4-5 respectively) school absence (chapter 7) and emergency presentation with self-harm (chapter 8), all of which are documented as part of routine clinical or educational practice. Because the

recorders (i.e. clinicians or school administrators) and participants were essentially blinded to group assignment (i.e the study hypotheses and analyses at the time of collection) the likelihood of the study findings being driven inadvertently by participants or observer biases were very small. This reduced the likelihood of non-random misclassification biases, which can lead to either overestimation of underestimation of exposure-outcome associations.

The NLP approaches described in this thesis, unlocked detailed clinical information held with free-text notes, including the nature of presenting symptoms or adverse effects. It also permitted comprehensive validation work to be conducted, to ensure that the structured information ascertained for analyses accurately represented clinical reality, such as the diagnostic assessment or treatment change. These data would not normally be available in an electronic case register derived purely from structured, administrative healthcare data.[65]

All the studies using data linkage approaches (chapters 6-8) were able to examine risk factors and outcomes, which are clinically relevant but not commonly documented within clinical environments, such as educational attainment or attendance. Data extracted from electronic health records are only as good as the information available within that individual system. [65] As described in chapter 8, this has been a limitation of self-harm outcome research using routinely collected clinical data. Whilst school factors are relevant determinants of self-harm risk in adolescents, very few health database studies have been able to capture school data, therefore it has not been included in analysis models. As illustrated within chapter 8, linkage studies can provide total local population samples within which clinical samples are imbedded. The study had follow-up data on very large samples, with participation and retention of many individuals at risk who traditionally may be lost to follow-up. Because the information available in the total population data (in this case the NPD or HES) was limited, the linkages to in-depth records held in CRIS, can reduce misclassification and provide better interpretation and analysis of the outcome of interest.[372]

9.3 LIMITATIONS

There are several limitations which need to be considered when drawing conclusions from the studies in this thesis. First, all of the clinical outcome and exposure data were derived from samples of people who have had contact with SLaM NHS health services or state-maintained

schools. Children who were not in primary or secondary state funded educational services, between Sept 2007 and 2013, for example in private school, or been taught at home, would have been excluded from the population sample, and not have educational data linked into the clinical data (chapters 6-8). In terms of clinical sampling, these data will not capture children and adolescents who met criteria for an ASD or Psychosis diagnosis but had no contact with SLaM clinical services over the data collection / extraction periods. This approach may have excluded children and adolescents who have been managed exclusively by the private or voluntary sector, or those resident within the SLaM catchment area, or received care through primary or non-SLaM NHS services. However, there are a number of reasons why these limitations are unlikely to have significantly impacted the results on this thesis. Given the current clinical guidelines on the use of antipsychotics in children and adolescents with ASD or psychosis, it is very unlikely these conditions would be exclusively managed within primary care.[103,183] In the absence of local or national health economic data,[373] it is difficult to estimate the extent to which private or voluntary services would manage children with ASD or psychosis without some contact with specialist services that provide care for the local catchment. SLaM provides the most comprehensive CAMHS in England and Wales, and few clinical presentations cannot be met by the diverse range of specialist inpatient outpatient services within SLaM.

The regional scope of the data could incur selection biases towards healthier samples. Children who move residence frequently are known to have higher rates of psychopathology,[374] so only having access to regional clinical data, as opposed national data, may have lost more mobile populations, and introduced bias into the study. That said, a bias in the other direction may have occurred where SLaM provided care to more severely disabled populations external to its catchment area. This may have reduced the representativeness of the clinical sample to the local community, and created spurious associations between risk factors and outcomes due to selection procedures.[375] When relevant to questions posed to the studies in this thesis, I attempted to reduce this bias by extracting samples either continuously resident within local catchment area (chapter 8) or used sensitivity analyses to examined the generalisability of the findings to the local sample (chapters 2, 4 and 5). Hence, it is unlikely that the samples in thesis are unrepresentative of the source population.

A related limitation includes the restriction of age to the clinical samples, so that all clinical outcomes occurred prior to age 18. One of the reasons I imposed this was to reduce the heterogeneity in clinical practice often experienced by children with long term conditions, such

as ASD and psychosis, when they become adults and move from CAMHS to adult psychiatric services.[376] I was concerned that this heterogeneity may have had considerable influence on the way clinical data is recorded, as well as the mental health treatments offered and outcomes obtained.[377] For example, CAMHS clinicians may be more inclined/ disinclined undertake a clinical assessment, and negate or affirm the presence of certain symptoms, risk factors or treatments than colleagues in adult services.[378] All the analyses using clinical data in this thesis, are limited by the assumption that the clinical constructs are recorded in the data (i.e. diagnoses, symptoms, risks of harm or types of event), are reliable, and have good construct validity.[379] I have assumed that the constructs used are consistent across the study samples. It also assumes that variation in clinical data quality across services does not systematically bias the outcomes under investigation. By restricting the sample to under 18's, I hoped to limit the variation in service provision, clinical expertise and experience, and as a consequence, reduce the variation in clinical note taking and treatment.

Another issue with CRIS data, or any electronic health record data, is that exposures and outcomes of interest are only recorded during an encounter with the health system. [371] The more information recorded, generally the greater likelihood of characterising the patient with the symptoms of interest. Restricting the duration of records available to collect variables of interest, can reduce this information bias, but it will not provide complete mitigation. For example, seven days of an adolescent's health record data which includes a 3000 word Mental Health Act Tribunal Report and inpatient observations entered every 6 hours, will contain a greater amount of clinical information than an individual who has, over the course of a week, had one phone call to rebook an appointment. In the case of the former, there is a much greater chance for exposure of interests, like negative symptoms, risk assessment or psychopathology scales to be recorded, hence there is a difference in risk of symptom detection. Also, in contrast to cohort studies, follow-up measures extracted from CRIS are not typically achieved on a fixed interval. Limited clinical information over the course of treatment may relate to a good recovery with limited clinical contact after 6 months or limited engagement with treatment and poor recovery. It is difficult to judge how these biases affect the results in this thesis; they could lead to both an underestimate or overestimate of effects on the outcome. Where possible I have attempted to mitigate these biases by ensuring that the outcome of interest can only occur if there is engagement with health services (i.e. antipsychotic treatment) or when looking at school based outcomes it is collected on all pupils (i.e. school attendance). Also, I have looked to test hypotheses which aim to determine whether the exposures of interest are risk factors rather than

protective factors. Taking findings in chapter 5 as an example, where negative symptoms were associated with a significant increase risk of multiple treatment failure. Given that negatives symptoms were likely to misclassified as not present, due to the sensitivity of the NLP tool and lack of clinical documentation, the size of effects reported are likely to be an underestimate.

Finally, as with all observational studies, residual confounding is a potential limitation. Any of the exposure – outcome associations reported may be potentially explained by unmeasured factors, and hence risk factors identified do not necessarily cause the outcomes reported.

9.4 IMPLICATIONS

The clinical, research and policy implications of this thesis can be derived, both from the study findings, and the methodologies employed. Regarding the former, the findings presented extend the knowledge base relating to neurodevelopmental comorbidity and clinical outcomes. In general, they provide empirical support for the hypothesis that neurodevelopmental comorbidities increase children and adolescents' risk for potentially more harmful treatments, greater treatment complexity and worse clinical outcomes.

These findings, I hope, will support the drive for intervention trials which include children with psychiatric co-morbidities, and discontinue the practice of stipulating psychiatric co-morbidity as exclusion criteria to maintain internal validity.[111] As illustrated in chapter 2, over 80% of children with ASD treated by antipsychotics had at least one psychiatric comorbidity. Many published trials examining the efficacy of antipsychotic use for managing challenging behaviours in ASD have excluded comorbid groups.[380] This demonstrates a profound mismatch between antipsychotic trial based samples and those patients who are most likely to receive the treatment in real world clinical settings. Related to this, the studies presented in chapter 4 and 5 which examine samples with early onset psychosis, highlight that it is possible to find markers for pharmacological treatment difficulties, at the early stages of pharmacological treatment. The findings demonstrate that recovery and response to antipsychotic treatment are especially problematic for a considerable proportion of young people. I found around one fifth of young people had tried three different antipsychotic treatments prior to turning 18, 30% of these had an insufficient response to their treatments. These studies show that identifying certain aspects of the clinical phenotype around first presentation, such as ASD and or negative symptoms, could help clinicians discern who is more vulnerable to a complex treatment path, and hence who may

benefit from early adjunctive therapeutic strategies. These data also reflect a need for robust evidence to be provided for CAMHS clinician's to determine if clozapine treatment is effective in early onset psychosis.[183] Without this evidence, a large proportion of proportion of children and adolescents will continue to be underserved by current therapeutic strategies in CAMHS.

The final study in chapter 8, provides robust evidence that ASD and ADHD, and a number of other population level educational factors, predict potentially severe self-harm in adolescence. These findings have implications for parents and teachers, as well as clinicians and public health policy. This is the first time ASD has been recognised as population level risk factor for self-harm; as far as I know, no studies that have looked at this association before. Increasing awareness that male children and adolescents with ASD are more likely to experience self-harm is an important first step in developing early recognition and future prevention programmes within schools and other child orientated services. Because this study is the first to identify ASD as a risk factor for self-harm, there will be a greater need for education and training tools to be developed, so that professionals can recognise and manage self-harm proficiently within ASD populations. The majority of children in this study, who presented with self-harm to emergency departments, had no recorded clinical contact with specialist mental health services. This does not imply these children's mental health needs were being neglected, but shows the high levels of morbidity which school services and families may have to manage without specialist NHS support. The findings reinforce the educational sectors role in monitoring vulnerable pupils, and, where resources are limited, the potential for repurposing existing routinely collected educational data to identify vulnerable groups.

The methodologies employed in the thesis have a number of implications too. The first relates to privacy. Information was collected with opt-out consent for health data, and, with no consent for education data. As explained in chapter 7, it is likely that gaining consent would scientifically invalidate the work through loss of representative samples, and, unless bolstered by significant resources, would not provide sufficient power to test the hypotheses proposed. The computing techniques and the governance methods used within this thesis, have enabled personal data to be collected and linked, with little risk of breeching confidentiality and privacy. Exposure to identifiable information has been further limited by the automated methods provided by the NLP approaches. This means that mainly computers, rather than human raters, are extracting information and securely storing information from personal health records.

In this thesis, I describe how these processes can deliver clinically relevant research findings, and offer very high levels of protection relating confidentiality and privacy. Whilst it is known the public can appreciate the benefits of this developing technology, it is counterbalanced by a lack of trust that large organisations will always use individual data responsibly. Recurrent stories of data security breaches and potential for misuse, provoke concerns about the use of health data beyond direct clinical use.[280] Personal health data are viewed as confidential, private and sensitive, and should not be shared outside secure, authorised bodies such as the NHS.[381] This thesis provides evidence that sharing very limited amounts of identifiable data for linkage to organisations outside of the NHS for population level analysis, is beneficial. I hope that the evidence in this thesis can be used to strengthen public engagement campaigns designed to highlight the positive impact of health informatics research on public health.

The opportunities and potential of NLP, as presented in this thesis, are extensive for health research. I show that NLP applied to records routinely collected by health service providers can accurately quantify an array of patient characteristics across emergency, community and inpatient settings. As extraction techniques improve and the data sources become more detailed, their potential for determining individual prognoses, treatment effectiveness and potential harms are all within greater reach, requiring far less resources than would be needed using primary data collection approaches studies. As shown in the thesis, NLP approaches enable psychiatric epidemiologists to improve risk factor and disease identification, and to build richer characterisations of child and adolescent clinical samples than can be achieved by the use of structured data alone.[382]

The CRIS system offers a sustainable resource for population-based analyses of linked patient level data and provides a powerful platform for continuous evaluation of local child health policy initiatives.[252] CRIS is being reproduced in other areas. Currently fourteen providers of mental healthcare in England, many of them covering child and adolescent mental health services, have developed systems based on the CRIS model.[383] The NLP and data linkage approaches employed in this thesis could be applied to these records in the other healthcare centres. However, local variations in clinician recording, and data storage and retrieval systems could be extensive, even across UK NHS Trusts. [384,385] Future work could involve testing the extent to which some of the NLP and linkage algorithms used in this thesis can be directly transposed on alternative health record systems.

9.5 FUTURE RESEARCH DIRECTIONS

Looking ahead, there are clearly substantial opportunities to improve mental healthcare using NLP and data linkage methodologies. As described in chapter 6, these data and the methods used to acquire them, have considerable potential for population-level analyses which can support public health interventions and policy evaluation, but the question remains whether these approaches can yet be applied to direct clinical care. For example, can NLP applications be used to analyse information from an individual's record to help direct clinical decisions, such as an appropriate medication choice? [70] Or can data linkages be used to inform other public services of a health event, such as a pupil's presentation to A&E with self-harm?

Before these applications are employed into direct clinical care, I believe there are a number of questions that need to be examined in future research work. The first relates to accuracy. As the work in chapter 3 demonstrates, the ability to detect accurate positive or negative affirmation of suicidality in health records is a complex task, even without introducing the concept of chronology, i.e. accurately detecting whether positive suicidal references are occurring presently or in the past. If clinicians want to use NLP to acquire a rapid and accurate synthesis of their patients' records, NLP techniques will need to be developed and evaluated, which can accurately identify the temporal aspects of suicidality, other psychopathological states and their treatments. Again, in relation to accuracy, clinicians will need to know the comparative validity, reliability and cost effectiveness of these NLP tools against 'clinician as usual' practices. Research will need be conducted on NLP outputs regarding clinical acceptability, according to their application. For example, if a clinician wants a complete medication history for one his patients with corresponding clinical response over time, would they be satisfied to make a clinical decision from a NLP-driven output? Does information, with 80% accuracy on drug type, timing and clinical improvement, which can be delivered within a minute, provide a clinician with sufficient confidence they will make an informed choice? Does it provide better performance relative to a thirty minute clinician note review? And if so, do these information systems improve efficiency and clinical outcomes to make them economically and ethically viable?

Regarding using data linkages for direct care across different services, I demonstrate in this thesis that without a shared unique identification number, there is a significant risk of mis-matching

between health and education service data. Also, this risk is not equally shared across age and demographic groups. In epidemiological studies, statistical adjustments can remove some of these biases. However, they cannot be applied to an individual record, and a missed or falsely matched record could have profound implications for direct clinical care and public trust.[386] Trials are needed to evaluate the acceptability and outcomes of processes that automatically link and disclose sensitive information to other public services. The capability to share and link information across services will have benefits, but may also harm.[387] Significant research and investment is needed to create and evaluate systems which provide the right mixture of automated and human rater decisions,[388] to support real time, near perfect matching and analysis of cross sector records. The computational power exists now to run linked public sector data systems. Embedding them into clinical care, if approached carefully, is an exciting prospect for enhancing individual and public health care.

9.6 CONCLUSION

The results of the investigations contained in this thesis demonstrate Big Data techniques, specifically NLP and data linkages of electronic health records, have a clear role in clinical epidemiological studies of child and adolescent mental health. These tools, combined with the continued digitisation of public service activity, can unlock huge and detailed data resources for population-based analysis. However, current approaches have deficiencies. Limitations in accuracy, construct validity, and restrictions in the data available, suggest these methods are unlikely to supplant primary data collection approaches in the near future. The work in this thesis represents some of the first studies to apply these Big Data techniques to questions related to child mental health, within a NHS environment. It provides a precedent for researchers, data scientists, clinicians and local decision makers who are looking to better understand what is happening to children within their own local community. Big data methods are crucial areas for research and development, which should be embraced as a method to understand and enhance child and adolescent mental health – particularly as current policies drive the ubiquity of the electronic record in public services. As the thesis shows, this resource, combined with linked data and NLP approaches, can help capture the 'hard to reach' in research samples. Over time, I hope these techniques will continue to be developed, and become an essential tool in child and adolescence public mental health, enabling us to better highlight and address the inequities faced by children and adolescences with mental health disorders.

REFERENCES

1 Institute for Health Transformation. Big Data In Healthcare. iHT^2 New York, USA, 2013 http://ihealthtran.com/big-data-in-healthcare (accessed May 14, 2017).

2 Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst* 2014; **2**: 3.

3 Dutcher J. What Is Big Data? 2014. https://datascience.berkeley.edu/what-is-big-data/ (accessed Sept 8, 2017).

4 Kumar KPK, Geethakumari G. Detecting misinformation in online social networks using cognitive psychology. *Hum-Centric Comput Inf Sci* 2014; **4**: 14.

5 Royal Society Science Policy Centre. Science as an open enterprise: Final report. Royal Society, 2012 https://royalsociety.org/policy/projects/science-public-enterprise/Report/ (accessed Sept 14, 2017).

6 Jackson R, Ball M, Patel R, Hayes R, Dobson R, Stewart R. Text Hunter: A User Friendly Tool for Extracting Generic Concepts from Free Text in Clinical Research. *Proc Am Med Inform Assoc* 2014; : 729–38.

7 Green H, McGinnity Á, Meltzer H, Ford T, Goodman R, others. Mental health of children and young people in Great Britain, 2004. Palgrave Macmillan Basingstoke, 2005.

8 McManus S, Bebbington P, Jenkins R, Brugha T, editors. Mental health and wellbeing in England: Adult Psychiatric Morbidity Survey 2014. Leeds: NHS Digital, 2016.

9 Department of Health. Chief Medical Officer annual report: public mental health. Department of Health, 2014 https://www.gov.uk/government/publications/chief-medical-officer-cmo-annual-report-public-mental-health (accessed Sept 9, 2017).

10 NICE. Mental health and behavioural conditions: Guidance and guideline topic. https://www.nice.org.uk/guidance/conditions-and-diseases/mental-health-and-behavioural-conditions (accessed June 23, 2017).

11 Public Health England. Atlas of Variation in mental health care. UK, 2011 http://fingertips.phe.org.uk/documents/Atlas_2011_MentalHealth.pdf (accessed Sept 9, 2017).

12 Wolfe I, Thompson M, Gill P, *et al.* Health services for children in western Europe. *The Lancet* 2013; **381**: 1224–34.

13 Kim-Cohen J, Caspi A, Moffitt TE, Harrington H, Milne BJ, Poulton R. Prior juvenile diagnoses in adults with mental disorder: developmental follow-back of a prospective-longitudinal cohort. *Arch Gen Psychiatry* 2003; **60**: 709–17.

14 Wilkinson RG, Marmot M. Social Determinants of Health: The Solid Facts. World Health Organisation Publishing. Copenhagen 2003. http://www.euro.who.int/__data/assets/pdf_file/0005/98438/e81384.pdf?ua=1 (accessed Sept 20,2017)

15 Suhrcke M, Pillas D, Selai C. Economic aspects of mental health in children and adolescents. In: Social cohesion for mental well-being among adolescents. World Health Organisation Publishing. Copenhagen, 2008. http://www.euro.who.int/__data/assets/pdf_file/0005/84623/E91921.pdf (accessed Sept 20, 2017)

16 OECD. Sick on the Job?: Myths and Realities about Mental Health and Work, OECD Publishing. Paris, 2012. http://dx.doi.org/10.1787/9789264124523-en (accessed Sept 20, 2017)

17 Fergusson DM, Horwood LJ, Ridder EM. Show me the child at seven: the consequences of conduct problems in childhood for psychosocial functioning in adulthood. *J Child Psychol Psychiatry* 2005; **46**: 837–49.

18 Kratzer L, Hodgins S. Adult outcomes of child conduct problems: a cohort study. *J Abnorm Child Psychol* 1997; **25**: 65–81.

19 Costello EJ. Early Detection and Prevention of Mental Health Problems: Developmental Epidemiology and Systems of Support. *J Clin Child Adolesc Psychol* 2016; **45**: 710–7.

20 Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 1996; **312**: 1215–8.

21 Caldwell PH, Murphy SB, Butow PN, Craig JC. Clinical trials in children. *Lancet* 2004; **364**: 803–11.

22 Rothwell PM. External validity of randomised controlled trials: 'to whom do the results of this trial apply?' *Lancet* 2005; **365**: 82–93.

23 Rutter M. Is Sure Start an Effective Preventive Intervention? *Child Adolesc Ment Health* 2006; **11**: 135–41.

24 Poulton R, Moffitt TE, Silva PA. The Dunedin Multidisciplinary Health and Development Study: overview of the first 40 years, with an eye to the future. *Soc Psychiatry Psychiatr Epidemiol* 2015; **50**: 679–93.

25 Melhuish E, Belsky J, Leyland AH, Barnes J. Effects of fully-established Sure Start Local Programmes on 3-year-old children and their families living in England: a quasi-experimental observational study. *Lancet* 2008; **372**: 1641–7.

26 Verhulst FC, Tiemeier H. Epidemiology of child psychopathology: major milestones. *Eur Child Adolesc Psychiatry* 2015; **24**: 607–17.

27 Burt A. Improving children and young people's mental health care. King's Fund, 2015 https://www.gov.uk/government/speeches/improving-children-and-young-peoples-mental-health-care (accessed June 23, 2017).

28 Karanikolos M, Mladovsky P, Cylus J, *et al.* Financial crisis, austerity, and health in Europe. *Lancet* 2013; **381**: 1323–31.

29 Association of teachers and lecturers. Mental Health Survey. 2015 https://www.atl.org.uk/Images/March%2026%20for%2028%202015%20-%20ATL%20Mental%20Health%20Survey.pdf (accessed July 15, 2017).

30 Livingstone S, Smith PK. Annual research review: Harms experienced by child users of online and mobile technologies: the nature, prevalence and management of sexual and aggressive risks in the digital age. *J Child Psychol Psychiatry* 2014; **55**: 635–54.

31 Knudsen AK, Hotopf M, Skogen JC, Overland S, Mykletun A. The health status of nonparticipants in a population-based health study: the Hordaland Health Study. *Am J Epidemiol* 2010; **172**: 1306–14.

32 Hatch SL, Woodhead C, Frissa S, *et al.* Importance of Thinking Locally for Mental Health: Data from Cross-Sectional Surveys Representing South East London and England. *PLoS ONE* 2012; **7**. doi:10.1371/journal.pone.0048012.

33 Helakorpi S, Mäkelä P, Holstila A, Uutela A, Vartiainen E. Can the accuracy of health behaviour surveys be improved by non-response follow-ups? *Eur J Public Health* 2015; **25**: 487–90.

34 Office for National Statistics. National Statistics Quality Review (NSQR) Series (2) Report Number 1: Review of the Labour Force Survey. ONS, 2014 http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/guide-method/method-quality/quality/quality-reviews/list-of-current-national-statistics-quality-reviews/nsqr-series--2--report-no--1/index.html (accessed June 20, 2017).

35 Office for National Statistics. Labour Force Survey performance and quality monitoring report. 2015 http://www.ons.gov.uk/ons/guide-method/method-quality/specific/labour-market/labour-force-survey/index.html.

36 Galea S, Tracy M. Participation rates in epidemiologic studies. *Ann Epidemiol* 2007; **17**: 643–53.

37 Wadman M. Child-study turmoil leaves bitter taste. *Nat News* 2012; **485**: 287.

38 Castles S, Haas H de, Miller MJ. The Age of Migration: International Population Movements in the Modern World. Palgrave Macmillan, London 2013.

39 Collins FS, Manolio TA. Merging and emerging cohorts: Necessary but not sufficient. *Nature* 2007; **445**: 259–259.

40 Cuthbert BN, Insel TR. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Med* 2013; **11**: 126.

41 Rani F, Murray M, Byrne P, Wong I. Epidemiologic features of antipsychotic prescribing to children and adolescents in primary care in the United Kingdom. *Paediatrics* 2008; **121**: 1002–9.

42 Edelsohn GA, Karpov I, Parthasarathy M, *et al.* Trends in Antipsychotic Prescribing in Medicaid-Eligible Youth. *J Am Acad Child Adolesc Psychiatry* 2017; **56**: 59–66.

43 Okkels N, Vernal DL, Jensen SOW, McGrath JJ, Nielsen RE. Changes in the diagnosed incidence of early onset schizophrenia over four decades. *Acta Psychiatr Scand* 2013; **127**: 62–8.

44 Kho ME, Duffett M, Willison DJ, Cook DJ, Brouwers MC. Written informed consent and selection bias in observational studies using medical records: systematic review. *BMJ* 2009; **338**: b866.

45 Wolke D, Waylen A, Samara M, *et al.* Selective drop-out in longitudinal studies and non-biased prediction of behaviour disorders. *Br J Psychiatry* 2009; **195**: 249–56.

46 Martin J, Tilling K, Hubbard L, *et al.* Association of Genetic Risk for Schizophrenia With Nonparticipation Over Time in a Population-Based Cohort Study. *Am J Epidemiol* 2016; **183**: 1149–58.

47 Doherty JL, Owen MJ. Genomic insights into the overlap between psychiatric disorders: implications for research and clinical practice. *Genome Med* 2014; **6**: 29.

48 Children & Young People's Mental Health Coalition. Overlooked and Forgotten. 2013 http://www.cypmhc.org.uk/resources/overlooked_and_forgotten_full_report/ (accessed Sept 14, 2017).

49 House of Commons Health Select Committee (HSC). Child and adolescent mental health and CAMHS Services—third report of session 2014–15. UK Government, 2014 http://www.publications.parliament.uk/pa/cm201415/cmselect/cmhealth/342/342.pdf (accessed Sept 14, 2017).

50 Ford DV, Jones KH, Verplancke J-P, *et al.* The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC Health Serv Res* 2009; **9**: 157.

51 SAIL Databank - The Secure Anonymised Information Linkage Databank. https://saildatabank.com/about-us/overview/ (accessed June 26, 2017).

52 Administrative Data Taskforce. UK Administrative Data Research Network: Improving Access for Research and Policy. Economic and Social Research Council, 2012 http://www.esrc.ac.uk/_images/ADT-Improving-Access-for-Research-and-Policy_tcm8-24462.pdf (accessed Sept 14, 2017).

53 Ford T, Edwards V, Sharkey S, *et al.* Supporting teachers and children in schools: the effectiveness and cost-effectiveness of the incredible years teacher classroom management programme in primary school children: a cluster randomised controlled trial, with parallel economic and process evaluations. *BMC Public Health* 2012; **12**: 719.

54 Holman CDJ, Bass AJ, Rosman DL, *et al.* A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system. *Aust Health Rev Publ Aust Hosp Assoc* 2008; **32**: 766–77.

55 Medical Research Council. MRC Review of mental health research: report of the Strategic Review Group. 2010 https://www.mrc.ac.uk/documents/pdf/mrc-review-of-mental-health-research-2010/ (accessed Aug 29, 2017).

56 Stewart R, Davis K. 'Big data' in mental health research: current status and emerging possibilities. *Soc Psychiatry Psychiatr Epidemiol* 2016; **51**: 1055–72.

57 Munk-Jørgensen P, Okkels N, Golberg D, Ruggeri M, Thornicroft G. Fifty years' development and future perspectives of psychiatric register research. *Acta Psychiatr Scand* 2014; **130**: 87–98.

58 Statistics Finland. Population census. 2015. http://www.stat.fi/tup/vl2010/index_en.html (accessed July 15, 2017).

59 Tromp M, Ravelli AC, Bonsel GJ, Hasman A, Reitsma JB. Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage. *J Clin Epidemiol* 2011; **64**: 565–72.

60 Harron K, Goldstein H, Dibben C. Methodological Developments in Data Linkage. Chichester, West Sussex, United Kingdom: Wiley-Blackwell, 2015.

61 Newcombe HB, Kennedy JM, Axford SJ, James AP. Automatic linkage of vital records. *Science* 1959; **130**: 954–9.

62 Brennan L, Watson M, Klaber R, Charles T. The importance of knowing context of hospital episode statistics when reconfiguring the NHS. *BMJ* 2012; **344**: e2432.

63 Morley KI, Wallace J, Denaxas SC, *et al.* Defining Disease Phenotypes Using National Linked Electronic Health Records: A Case Study of Atrial Fibrillation. *PLoS ONE* 2014; **9**. doi:10.1371/journal.pone.0110900.

64 Hagger-Johnson G, Harron K, Gonzalez-Izquierdo A, *et al.* Identifying possible false matches in anonymized hospital administrative data without patient identifiers. *Health Serv Res* 2015; **50**: 1162–78.

65 Stewart R. The big case register. *Acta Psychiatr Scand* 2014; **130**: 83–6.

66 NHS Digital. Hospital Episode Statistics. 2013; published online May 28. http://www.hscic.gov.uk/hes (accessed March 17, 2016).

67 Davis M. West Riding Pauper Lunatic Asylum Through Time.: Amberley Publishing Limited, Gloucestershire, UK 2014.

68 Perera G, Soremekun M, Breen G, Stewart R. The psychiatric case register: noble past, challenging present, but exciting future. *Br J Psychiatry* 2009; **195**: 191–3.

69 Gearing RE, Mian IA, Barber J, Ickowicz A. A methodology for conducting retrospective chart review research in child and adolescent psychiatry. *J Can Acad Child Adolesc Psychiatry* 2006; **15**: 126–34.

70 Wachter R. The Digital Doctor: Hope, Hype, and Harm at the Dawn of Medicine's Computer Age. : McGraw-Hill Education.New York, 2015.

71 McGuffin P, Farmer A, Harvey I. A polydiagnostic application of operational criteria in studies of psychotic illness. Development and reliability of the OPCRIT system. *Arch Gen Psychiatry* 1991; **48**: 764–70.

72 Rucker J, Newman S, Gray J, *et al.* OPCRIT+: an electronic system for psychiatric diagnosis and data collection in clinical and research settings. *Br J Psychiatry* 2011; **199**: 151–5.

73 NHS England. Digital Maturity Assessment 2015/2016. https://data.england.nhs.uk/dataset/digital-maturity-assessment-2015-2016 (accessed July 18, 2017).

74 Carter P, Laurie GT, Dixon-Woods M. The social licence for research: why care.data ran into trouble. *J Med Ethics* 2015; **41**: 404–9.

75 Roberts A. Language, Structure, and Reuse in the Electronic Health Record. *AMA J Ethics* 2017; **19**: 281.

76 Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform* 2008; **1**: 128–44.

77 Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. *J Am Med Inform Assoc JAMIA* 2011; **18**: 181–6.

78 Greenhalgh T, Potts HWW, Wong G, Bark P, Swinglehurst D. Tensions and paradoxes in electronic patient record research: a systematic literature review using the meta-narrative method. *Milbank Q* 2009; **87**: 729–88.

79 Chowdhury GG. Natural language processing. *Annu Rev Inf Sci Technol* 2003; **37**: 51–89.

80 Manning C. Foundations of Statistical Natural Language Processing, 2001 edition. Cambridge, Mass: MIT Press, 1999.

81 Uzuner Ö, Luo Y, Szolovits P. Evaluating the State-of-the-Art in Automatic De-identification. *J Am Med Inform Assoc* 2007; **14**: 550–63.

82 Lingren T, Chen P, Bochenek J, *et al.* Electronic Health Record Based Algorithm to Identify Patients with Autism Spectrum Disorder. *PloS One* 2016; **11**: e0159621.

83 Murphy SN, Weber G, Mendis M, *et al.* Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc* 2010; **17**: 124–30.

84 Sohn S, Kocher J-PA, Chute CG, Savova GK. Drug side effect extraction from clinical narratives of psychiatry and psychology patients. *J Am Med Inform Assoc JAMIA* 2011; **18 Suppl 1**: i144–9.

85 Murff HJ, FitzHenry F, Matheny ME, *et al.* Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA* 2011; **306**: 848–55.

86 Perlis RH, Iosifescu DV, Castro VM, *et al.* Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. *Psychol Med* 2012; **42**: 41–50.

87 Wolters Kluwer Health. OVID Gateway. http://gateway.ovid.com.

88 O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Syst Rev* 2015; **4**: 5.

89 Clements CC, Castro VM, Blumenthal SR, *et al.* Prenatal antidepressant exposure is associated with risk for attention-deficit hyperactivity disorder but not autism spectrum disorder in a large health system. *Mol Psychiatry* 2015; **20**: 727–34.

90 Castro VM, Kong SW, Clements CC, *et al.* Absence of evidence for increase in risk for autism or attention-deficit hyperactivity disorder following antidepressant exposure during pregnancy: a replication study. *Transl Psychiatry* 2016; **6**: e708.

91 Doshi-Velez F, Ge Y, Kohane I. Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics* 2014; **133**: e54–63.

92 Kohane IS, McMurry A, Weber G, *et al.* The co-morbidity burden of children and young adults with autism spectrum disorders. *PloS One* 2012; **7**: e33224.

93 Lyalina S, Percha B, LePendu P, Iyer SV, Altman RB, Shah NH. Identifying phenotypic signatures of neuropsychiatric disorders from electronic medical records. *J Am Med Inform Assoc* 2013; **20**: e297–305.

94 Anderson HD, Pace WD, Brandt E, *et al.* Monitoring suicidal patients in primary care using electronic health records. *J Am Board Fam Med* 2015; **28**: 65–71.

95 Metzger M-H, Tvardik N, Gicquel Q, Bouvry C, Poulet E, Potinet-Pagliaroli V. Use of emergency department electronic medical records for automated epidemiological surveillance of suicide attempts: a French pilot study. *Int J Methods Psychiatr Res* 2017; **26**: e1522.

96 McIntosh AM, Stewart R, John A, *et al.* Data science for mental health: a UK perspective on a global challenge. *Lancet Psychiatry* 2016; **3**: 993–8.

97 Poulin C, Shiner B, Thompson P, *et al.* Predicting the Risk of Suicide by Analyzing the Text of Clinical Notes. *PLoS ONE* 2014; **9**: e85733.

98 Frost J, Okun S, Vaughan T, Heywood J, Wicks P. Patient-reported Outcomes as a Source of Evidence in Off-Label Prescribing: Analysis of Data From PatientsLikeMe. *J Med Internet Res* 2011; **13**. doi:10.2196/jmir.1643.

99 Ennis L, Robotham D, Denis M, *et al.* Collaborative development of an electronic Personal Health Record for people with severe and enduring mental health problems. *BMC Psychiatry* 2014; **14**. doi:10.1186/s12888-014-0305-9.

100 Young Z, Craven MP, Groom M, Crowe J. Snappy App: A Mobile Continuous Performance Test with Physical Activity Measurement for Assessing Attention Deficit

Hyperactivity Disorder. In: Kurosu M, ed. Human-Computer Interaction. Applications and Services. Springer International Publishing, 2014: 363–73.

101 Arora S, Venkataraman V, Zhan A, *et al.* Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study. *Parkinsonism Relat Disord* 2015; **21**: 650–3.

102 Harpaz R, DuMouchel W, Shah NH, Madigan D, Ryan P, Friedman C. Novel Data Mining Methodologies for Adverse Drug Event Discovery and Analysis. *Clin Pharmacol Ther* 2012; **91**: 1010–21.

103 NICE. Autism: the management and support of children and young people on the autism spectrum. CG 170. National Institute for Health and Care Excellence. London. 2013.

104 Baghdadli A, Pascal C, Grisi S, Aussilloux C. Risk factors for self-injurious behaviours among 222 young children with autistic disorders. *J Intellect Disabil Res* 2003; **47**: 622–7.

105 Hsia Y, Wong AYS, Murphy DGM, Simonoff E, Buitelaar JK, Wong ICK. Psychopharmacological prescriptions for people with autism spectrum disorder (ASD): a multinational study. *Psychopharmacology (Berl)* 2013; **231**: 999–1009.

106 Coury DL, Anagnostou E, Manning-Courtney P, *et al.* Use of Psychotropic Medication in Children and Adolescents With Autism Spectrum Disorders. *Pediatrics* 2012; **130**: S69–76.

107 Mandell DS, Morales KH, Marcus SC, Stahmer AC, Doshi J, Polsky DE. Psychotropic Medication Use Among Medicaid-Enrolled Children With Autism Spectrum Disorders. *Pediatrics* 2008; **121**: e441–8.

108 Murray ML, Hsia Y, Glaser K, *et al.* Pharmacological treatments prescribed to people with autism spectrum disorder (ASD) in primary health care. *Psychopharmacology (Berl)* 2013; **231**: 1011–21.

109 Bachmann CJ, Manthey T, Kamp-Becker I, Glaeske G, Hoffmann F. Psychopharmacological treatment in children and adolescents with autism spectrum disorders in Germany. *Res Dev Disabil* 2013; **34**: 2551–63.

110 McPheeters ML, Warren Z, Sathe N, *et al.* A Systematic Review of Medical Treatments for Children With Autism Spectrum Disorders. *Pediatrics* 2011; **127**: e1312–21.

111 American Academy of Child & Adolescent Psychiatry. Policy Statement on Comorbidity Treatment in Autism Spectrum Disorders and Intellectual Disabilities. http://www.aacap.org/aacap/Policy_Statements/2013/Comorbidity_Treatment_in_Autism_Spectrum_Disorders_and_Intellectual_Disabilities.aspx (accessed July 12, 2017).

112 Stringaris A. Irritability in children and adolescents: a challenge for DSM-5. *Eur Child Adolesc Psychiatry* 2011; **20**: 61–6.

113 NICE. Antisocial behaviour and conduct disorders in children and young people. QS59. National Institute for Health and Care Excellence. London 2014.

114 Dove D, Warren Z, McPheeters ML, Taylor JL, Sathe NA, Veenstra-VanderWeele J. Medications for Adolescents and Young Adults With Autism Spectrum Disorders: A Systematic Review. *Pediatrics* 2012; **130**: 717–26.

115 Lin J-D. Medical Care Burden of Children with Autism Spectrum Disorders. *Rev J Autism Dev Disord* 2014; **1**: 242–7.

116 Almandil NB, Liu Y, Murray ML, Besag FMC, Aitchison KJ, Wong ICK. Weight gain and other metabolic adverse effects associated with atypical antipsychotic treatment of children and adolescents: a systematic review and meta-analysis. *Paediatr Drugs* 2013; **15**: 139–50.

117 Glover G, Bernard S, Branford D, Holland A, Strydom A. Use of medication for challenging behaviour in people with intellectual disability. *Br J Psychiatry* 2014; **205**: 6–7.

118 Simonoff E, Pickles A, Charman T, Chandler S, Loucas T, Baird G. Psychiatric Disorders in Children With Autism Spectrum Disorders: Prevalence, Comorbidity, and Associated Factors in a Population-Derived Sample. *J Am Acad Child Adolesc Psychiatry* 2008; **47**: 921–9.

119 Every-Palmer S, Howick J. How evidence-based medicine is failing due to biased trials and selective publication. *J Eval Clin Pract* 2014; **20**: 908–14.

120 US Department of Health and Human Services. Aripiprazole: Pediatric Labeling Information. http://www.accessdata.fda.gov/scripts/cder/drugsatfda/index.cfm?fuseaction=Search.Overview&DrugName=ABILIFY8 (accessed Sept 10, 2017).

121 US Department of Health and Human Services. Risperidone : Pediatric Labeling Information. http://www.accessdata.fda.gov/scripts/cder/drugsatfda/index.cfm?fuseaction=Search.Overview&DrugName=RISPERIDONE (accessed Sept 10, 2017).

122 Taylor E. Pediatric psychopharmacology: Too much and too little. *World Psychiatry* 2013; **12**: 124–5.

123 World Health Organisation. Multiaxial Classification Child And Adolescent Psychiatric Disorders: The ICD-10 Classification of Mental and Behavioural Disorders in Children and Adolescents. World Health Organisation Publishing. Cambridge: 2008.

124 Lord C, Rutter M, Goode S, *et al.* Autism diagnostic observation schedule: A standardized observation of communicative and social behavior. *J Autism Dev Disord* 1989; **19**: 185–212.

125 Baird G, Simonoff E, Pickles A, *et al.* Prevalence of disorders of the autism spectrum in a population cohort of children in South Thames: the Special Needs and Autism Project (SNAP). *Lancet* 2006; **368**: 210–5.

126 Goodman R. The Strengths and Difficulties Questionnaire: a research note. *J Child Psychol Psychiatry* 1997; **38**: 581–6.

127 Stewart R, Soremekun M, Perera G, *et al.* The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register: development and descriptive data. *BMC Psychiatry* 2009; **9**: 51.

128 Fernandes AC, Cloete D, Broadbent MT, *et al.* Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records. *BMC Med Inform Decis Mak* 2013; **13**: 71.

129 Perera G, Broadbent M, Callard F, *et al.* Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource. *BMJ Open* 2016; **6**: e008721.

130 Cunningham H. GATE, a General Architecture for Text Engineering. *Comput Humanit* 2002; **36**: 223–54.

131 Cunningham H, Tablan V, Roberts A, Bontcheva K. Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS Comput Biol* 2013; **9**: e1002854.

132 Hayes RD, Downs J, Chang C-K, *et al.* The Effect of Clozapine on Premature Mortality: An Assessment of Clinical Monitoring and Other Potential Confounders. *Schizophr Bull* 2015; **41**: 644–55.

133 Shaffer D, Gould MS, Brasic J, *et al.* A Children's Global Assessment Scale (CGAS). *Arch Gen Psychiatry* 1983; **40**: 1228–31.

134 Wagner A, Lecavalier L, Arnold LE, Aman MG. Developmental disabilities modification of the Children's Global Assessment Scale. *Biol Psychiatry* 2007; **61**: 504–11.

135 Department for Communities and Local Government. English indices of deprivation 2010: technical report. https://www.gov.uk/government/statistics/english-indices-of-deprivation-2010-technical-report (accessed July 10, 2017).

136 Goodman R. Psychometric properties of the strengths and difficulties questionnaire. *J Am Acad Child Adolesc Psychiatry* 2001; **40**: 1337–45.

137 Rubin DM, Feudtner C, Localio R, Mandell DS. State Variation in Psychotropic Medication Use by Foster Care Children With Autism Spectrum Disorder. *Pediatrics* 2009; **124**: e305–12.

138 Garland AF, Brookman-Frazee L, Gray E. The Role of Parent Characteristics in Community-Based Medication Treatment for Children with Disruptive Behavior Problems. *Community Ment Health J* 2013; **49**: 507–14.

139 Leslie LK, Weckerly J, Landsverk J, Hough RL. Racial/ethnic differences in the use of psychotropic medication in high-risk children and adolescents. *J Am Acad Child Adolesc Psychiatry* 2003; **42**: 1433–42.

140 Zuvekas SH, Vitiello B, Norquist GS. Recent trends in stimulant medication use among U.S. children. *Am J Psychiatry* 2006; **163**: 579–85.

141 Drilea SK, Jowers K, Lichtenstein C, Hale M, Blau G, Stromberg S. Psychotropic medication use and clinical outcomes among children and adolescents receiving system of care services. *J Child Adolesc Psychopharmacol* 2013; **23**: 36–43.

142 Frazier TW, Shattuck PT, Narendorf SC, Cooper BP, Wagner M, Spitznagel EL. Prevalence and Correlates of Psychotropic Medication Use in Adolescents with an Autism Spectrum Disorder with and without Caregiver-Reported Attention-Deficit/Hyperactivity Disorder. *J Child Adolesc Psychopharmacol* 2011; **21**: 571–9.

143 Aman MG, Kasper W, Manos G. Line-item analysis of the Aberrant Behavior Checklist: results from two studies of aripiprazole in the treatment of irritability associated with autistic disorder. *J Child Adolesc Psychopharmacol* 2010; **20**: 415–22.

144 Hawley CJ, Littlechild B, Sivakumaran T. Structure and content of risk assessment proformas in mental healthcare. *J Ment Health* 2006; **15**: 437–48.

145 Waris P, Lindberg N, Kettunen K, Tani P. The relationship between Asperger's syndrome and schizophrenia in adolescence. *Eur Child Adolesc Psychiatry* 2013; **22**: 217–23.

146 Kyriakopoulos M, Stringaris A, Manolesou S, *et al.* Determination of psychosis-related clinical profiles in children with autism spectrum disorders using latent class analysis. *Eur Child Adolesc Psychiatry* 2015; **24**: 301–7.

147 Kumra S, Jacobsen LK, Lenane M, *et al.* 'Multidimensionally Impaired Disorder': Is It a Variant of Very Early-Onset Schizophrenia? *J Am Acad Child Adolesc Psychiatry* 1998; **37**: 91–9.

148 Stayer C, Sporn A, Gogtay N, *et al.* Multidimensionally impaired: the good news. *J Child Adolesc Psychopharmacol* 2005; **15**: 510–9.

149 McDougle CJ, Scahill L, Aman MG, *et al.* Risperidone for the core symptom domains of autism: results from the study by the autism network of the research units on pediatric psychopharmacology. *Am J Psychiatry* 2005; **162**: 1142–8.

150 Mayes SD, Calhoun SL, Baweja R, Mahr F. Suicide ideation and attempts in children with psychiatric disorders and typical development. *Crisis J Crisis Interv Suicide Prev* 2015; **36**: 55–60.

151 Segers M, Rawana J. What Do We Know About Suicidality in Autism Spectrum Disorders? A Systematic Review. *Autism Res* 2014; **7**: 507–21.

152 Burack JA, Iarocci G, Flanagan TD, Bowler DM. On mosaics and melting pots: conceptual considerations of comparison and matching strategies. *J Autism Dev Disord* 2004; **34**: 65–73.

153 Caspi A, Langley K, Milne B, *et al.* A replicated molecular genetic basis for subtyping antisocial behavior in children with attention-deficit/hyperactivity disorder. *Arch Gen Psychiatry* 2008; **65**: 203–10.

154 Raja M. Suicide risk in adults with Asperger's syndrome. *Lancet Psychiatry* 2014; **1**: 99–101.

155 Barak-Corren Y, Castro VM, Javitt S, *et al.* Predicting Suicidal Behavior From Longitudinal Electronic Health Records. *Am J Psychiatry* 2016; **174**: 154–62.

156 Downs J, Gilbert R, Hayes RD, Hotopf M, Ford T. Linking health and education data to plan and evaluate services for children. *Arch Dis Child* 2017; **102**: 599–602.

157 Haerian K, Salmasian H, Friedman C. Methods for identifying suicide or suicidal ideation in EHRs. *AMIA Annu Symp Proc AMIA Symp* 2012; **2012**: 1244–53.

158 Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc* 1994; **1**: 161–74.

159 Gkotsis G, Velupillai S, Oellrich A, Dean H, Liakata M, Dutta R. Don't Let Notes Be Misunderstood: A Negation Detection Method for Assessing Risk of Suicide in Mental Health Records. In: The Third Computational Linguistics and Clinical Psychology Workshop (CLPsych). 2016: 95–105.

160 Downs J, Hotopf M, Ford T, *et al.* Clinical predictors of antipsychotic use in children and adolescents with autism spectrum disorders: a historical open cohort study using electronic health records. *Eur Child Adolesc Psychiatry* 2016; **25**: 649–58.

161 Ting SA, Sullivan AF, Boudreaux ED, Miller I, Camargo CA. Trends in US emergency department visits for attempted suicide and self-inflicted injury, 1993–2008. *Gen Hosp Psychiatry* 2012; **34**: 557–65.

162 Richa S, Fahed M, Khoury E, Mishara B. Suicide in Autism Spectrum Disorders. *Arch Suicide Res* 2014; **18**: 327–39.

163 Harkema H, Dowling JN, Thornblade T, Chapman WW. Context: An Algorithm for Determining Negation, Experiencer, and Temporal Status from Clinical Reports. *J Biomed Inform* 2009; **42**: 839–51.

164 Schimmelmann BG, Conus P, Cotton S, McGorry PD, Lambert M. Pre-treatment, baseline, and outcome differences between early-onset and adult-onset psychosis in an epidemiological cohort of 636 first-episode patients. *Schizophr Res* 2007; **95**: 1–8.

165 Ballageer T, Malla A, Manchanda R, Takhar J, Haricharan R. Is adolescent-onset first-episode psychosis different from adult onset? *J Am Acad Child Adolesc Psychiatry* 2005; **44**: 782–9.

166 Hui CL-M, Li AW-Y, Leung C-M, *et al.* Comparing illness presentation, treatment and functioning between patients with adolescent- and adult-onset psychosis. *Psychiatry Res* 2014; **220**: 797–802.

167 Rajji TK, Ismail Z, Mulsant BH. Age at onset and cognition in schizophrenia: Meta-analysis. *Br J Psychiatry* 2009; **195**: 286–93.

168 Stentebjerg-Olesen M, Pagsberg AK, Fink-Jensen A, Correll CU, Jeppesen P. Clinical Characteristics and Predictors of Outcome of Schizophrenia-Spectrum Psychosis in Children and Adolescents: A Systematic Review. *J Child Adolesc Psychopharmacol* 2016; **26**: 410–27.

169 Díaz-Caneja CM, Pina-Camacho L, Rodríguez-Quiroga A, Fraguas D, Parellada M, Arango C. Predictors of outcome in early-onset psychosis: a systematic review. *Npj Schizophr* 2015; **1**. doi:10.1038/npjschz.2014.5.

170 Cannon-Spoor HE, Potkin SG, Wyatt RJ. Measurement of Premorbid Adjustment in Chronic Schizophrenia. *Schizophr Bull* 1982; **8**: 470–84.

171 Del Rey-Mejías Á, Fraguas D, Díaz-Caneja CM, *et al.* Functional deterioration from the premorbid period to 2 years after the first episode of psychosis in early-onset psychosis. *Eur Child Adolesc Psychiatry* 2015; **24**: 1447–59.

172 Sporn AL, Addington AM, Gogtay N, *et al.* Pervasive developmental disorder and childhood-onset schizophrenia: comorbid disorder or a phenotypic variant of a very early onset illness? *Biol Psychiatry* 2004; **55**: 989–94.

173 Larson FV, Wagner AP, Jones PB, *et al.* Psychosis in autism: comparison of the features of both conditions in a dually affected cohort. *Br J Psychiatry* 2016; bjp.bp.116.187682.

174 Khandaker GM, Stochl J, Zammit S, Lewis G, Jones PB. A population-based longitudinal study of childhood neurodevelopmental disorders, IQ and subsequent risk of psychotic experiences in adolescence. *Psychol Med* 2014; **44**: 3229–38.

175 Pina-Camacho L, Parellada M, Kyriakopoulos M. Autism spectrum disorder and schizophrenia: boundaries and uncertainties. *BJPsych Adv* 2016; **22**: 316–24.

176 Hassan GAM, Taha GRA. Long term functioning in early onset psychosis: two years prospective follow-up study. *Behav Brain Funct BBF* 2011; **7**: 28.

177 Fleischhaker C, Schulz E, Tepper K, Martin M, Hennighausen K, Remschmidt H. Long-term course of adolescent schizophrenia. *Schizophr Bull* 2005; **31**: 769–80.

178 Starling J, Dossetor D. Pervasive developmental disorders and psychosis. *Curr Psychiatry Rep* 2009; **11**: 190–6.

179 Sheitman BB, Kraus JE, Bodfish JW, Carmel H. Are the negative symptoms of schizophrenia consistent with an autistic spectrum illness? *Schizophr Res* 2004; **69**: 119–20.

180 Frank J, Lang M, Witt SH, *et al.* Identification of increased genetic risk scores for schizophrenia in treatment-resistant patients. *Mol Psychiatry* 2015; **20**: 150–1.

181 Politte LC, Henry CA, McDougle CJ. Psychopharmacological interventions in autism spectrum disorder. *Harv Rev Psychiatry* 2014; **22**: 76–92.

182 Rapoport J, Chavez A, Greenstein D, Addington A, Gogtay N. Autism spectrum disorders and childhood-onset schizophrenia: clinical and biological contributions to a relation revisited. *J Am Acad Child Adolesc Psychiatry* 2009; **48**: 10–8.

183 NICE. Psychosis and Schizophrenia in Children and Young People: Recognition and Management (CG155). National Institute for Health and Care Excellence. London 2013 https://www.nice.org.uk/guidance/cg155 (accessed April 14, 2017).

184 Correll CU, Kishimoto T, Nielsen J, Kane JM. Quantifying Clinical Relevance in the Treatment of Schizophrenia. *Clin Ther* 2011; **33**: B16–39.

185 Suzuki T, Remington G, Mulsant BH, *et al.* Defining treatment-resistant schizophrenia and response to antipsychotics: A review and recommendation. *Psychiatry Res* 2012; **197**: 1–6.

186 Schneider C, Papachristou E, Wimberley T, *et al.* Clozapine use in childhood and adolescent schizophrenia: A nationwide population-based study. *Eur Neuropsychopharmacol* 2015; **25**: 857–63.

187 Lieberman JA, Stroup TS, McEvoy JP, *et al.* Effectiveness of Antipsychotic Drugs in Patients with Chronic Schizophrenia. *N Engl J Med* 2005; **353**: 1209–23.

188 Patel R, Wilson R, Jackson R, *et al.* Association of cannabis use with hospital admission and antipsychotic treatment failure in first episode psychosis: an observational study. *BMJ Open* 2016; **6**: e009888.

189 Kahn RS, Fleischhacker WW, Boter H, *et al.* Effectiveness of antipsychotic drugs in first-episode schizophrenia and schizophreniform disorder: an open randomised clinical trial. *Lancet* 2008; **371**: 1085–97.

190 Weiden PJ. Discontinuing and switching antipsychotic medications: understanding the CATIE schizophrenia trial. *J Clin Psychiatry* 2007; **68 Suppl 1**: 12–9.

191 Lord C, Rutter M, DiLavore P, Risi S. Autism Diagnostic Observation Schedule (ADOS). Western Psychological Services. Los Angeles, California. 2000.

192 Volkmar F, Siegel M, Woodbury-Smith M, *et al.* Practice parameter for the assessment and treatment of children and adolescents with autism spectrum disorder. *J Am Acad Child Adolesc Psychiatry* 2014; **53**: 237–57.

193 Lord C, Rutter M, Le Couteur A. Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J Autism Dev Disord* 1994; **24**: 659–85.

194 McLennan D, Barnes H, Noble M, Davies J, Garratt E, Dibben C. The English indices of deprivation 2010. HMG Department of Communities. London 2011.

195 Rabinowitz J, Harvey PD, Eerdekens M, Davidson M. Premorbid functioning and treatment response in recent-onset schizophrenia. *Br J Psychiatry J Ment Sci* 2006; **189**: 31–5.

196 Rabinowitz J, Napryeyenko O, Burba B, *et al.* Premorbid functioning and treatment response in recent-onset schizophrenia: prospective study with risperidone long-acting injectable. *J Clin Psychopharmacol* 2011; **31**: 75–81.

197 Research Units on Pediatric Psychopharmacology Autism Network. Randomized, controlled, crossover trial of methylphenidate in pervasive developmental disorders with hyperactivity. *Arch Gen Psychiatry* 2005; **62**: 1266–74.

198 Kästner A, Begemann M, Michel TM, *et al.* Autism beyond diagnostic categories: characterization of autistic phenotypes in schizophrenia. *BMC Psychiatry* 2015; **15**: 115.

199 Demjaha A, Egerton A, Murray RM, *et al.* Antipsychotic treatment resistance in schizophrenia associated with elevated glutamate levels but normal dopamine function. *Biol Psychiatry* 2014; **75**: e11–3.

200 Lin K-M, Poland R. Ethnicity, culture, and psychopharmacology. In: Psychopharmacology. The Fourth Generation of Progress. American College of Neuropsycopharmacology, Raven Press, New York: 2000: 1907–17.

201 McCutcheon R, Beck K, Bloomfield MAP, Marques TR, Rogdaki M, Howes OD. Treatment resistant or resistant to treatment? Antipsychotic plasma levels in patients with poorly controlled psychotic symptoms. *J Psychopharmacol Oxf Engl* 2015; **29**: 892–7.

202 Meng H, Schimmelmann BG, Mohler B, *et al.* Pretreatment social functioning predicts 1-year outcome in early onset psychosis. *Acta Psychiatr Scand* 2006; **114**: 249–56.

203 Wimberley T, Støvring H, Sørensen HJ, Horsdal HT, MacCabe JH, Gasse C. Predictors of treatment resistance in patients with schizophrenia: a population-based cohort study. *Lancet Psychiatry* 2016; **3**: 358–66.

204 Findling RL, Johnson JL, McClellan J, *et al.* Double-blind maintenance safety and effectiveness findings from the Treatment of Early-Onset Schizophrenia Spectrum (TEOSS) study. *J Am Acad Child Adolesc Psychiatry* 2010; **49**: 583–94; quiz 632.

205 Buitelaar JK, van der Gaag RJ. Diagnostic Rules for Children with PDD-NOS and Multiple Complex Developmental Disorder. *J Child Psychol Psychiatry* 1998; **39**: 911–9.

206 Remschmidt H, Theisen F. Early-onset schizophrenia. *Neuropsychobiology* 2012; **66**: 63–9.

207 Kane JM, Correll CU. Past and present progress in the pharmacologic treatment of schizophrenia. *J Clin Psychiatry* 2010; **71**: 1115–24.

208 Strauss J, Carpenter W, Bartko J. The diagnosis and understanding of schizophrenia. Part III. Speculations on the processes that underlie schizophrenic symptoms and signs. *Schizophr Bull* 1974; **11**: 61–9.

209 Tandon R, Nasrallah H, Keshavan M. Schizophrenia, 'just the facts' 4. Clinical features and conceptualization. *Schizophr Res* 2009; **110**: 1–23.

210 Bobes J, Arango C, Garcia-Garcia M, Rejas J, CLAMORS Study Collaborative Group. Prevalence of negative symptoms in outpatients with schizophrenia spectrum disorders treated with antipsychotics in routine clinical practice: findings from the CLAMORS study. *J Clin Psychiatry* 2010; **71**: 280–6.

211 Patel R, Jayatilleke N, Broadbent M, *et al.* Negative symptoms in schizophrenia: a study in a large clinical sample of patients using a novel automated method. *BMJ Open* 2015; **5**: e007619.

212 Álvarez-Jiménez M, Gleeson JF, Henry LP, *et al.* Road to full recovery: longitudinal relationship between symptomatic remission and psychosocial recovery in first-episode psychosis over 7.5 years. *Psychol Med* 2012; **42**: 595–606.

213 Austin SF, Mors O, Secher RG, *et al.* Predictors of recovery in first episode psychosis: the OPUS cohort at 10 year follow-up. *Schizophr Res* 2013; **150**: 163–8.

214 Milev P, Ho B-C, Arndt S, Andreasen NC. Predictive values of neurocognition and negative symptoms on functional outcome in schizophrenia: a longitudinal first-episode study with 7-year follow-up. *Am J Psychiatry* 2005; **162**: 495–506.

215 White C, Stirling J, Hopkins R, *et al.* Predictors of 10-year outcome of first-episode psychosis. *Psychol Med* 2009; **39**: 1447–56.

216 Kirkpatrick B, Buchanan RW, Ross DE, Carpenter WT. A separate disease within the syndrome of schizophrenia. *Arch Gen Psychiatry* 2001; **58**: 165–71.

217 Parellada M, Castro-Fornieles J, Gonzalez-Pinto A, *et al.* Predictors of functional and clinical outcome in early-onset first-episode psychosis: the child and adolescent first episode of psychosis (CAFEPS) study. *J Clin Psychiatry* 2015; **76**: e1441–8.

218 Strauss GP, Allen DN, Miski P, Buchanan RW, Kirkpatrick B, Carpenter WT. Differential patterns of premorbid social and academic deterioration in deficit and nondeficit schizophrenia. *Schizophr Res* 2012; **135**: 134–8.

219 Rapado-Castro M, Soutullo C, Fraguas D, *et al.* Predominance of symptoms over time in early-onset psychosis: a principal component factor analysis of the Positive and Negative Syndrome Scale. *J Clin Psychiatry* 2010; **71**: 327–37.

220 Chang WC, Tang JYM, Hui CLM, *et al.* The relationship of early premorbid adjustment with negative symptoms and cognitive functions in first-episode schizophrenia: a prospective three-year follow-up study. *Psychiatry Res* 2013; **209**: 353–60.

221 Vyas NS, Hadjulis M, Vourdas A, Byrne P, Frangou S. The Maudsley early onset schizophrenia study. Predictors of psychosocial outcome at 4-year follow-up. *Eur Child Adolesc Psychiatry* 2007; **16**: 465–70.

222 Downs J, Lechler S, Dean H, *et al.* The association between co-morbid autism spectrum disorders and antipsychotic treatment failure in early-onset psychosis: a historical cohort study using electronic health records. *J Clin Psychiatry* 2017; **(in press)**.

223 Laruelle M. Schizophrenia: from dopaminergic to glutamatergic interventions. *Curr Opin Pharmacol* 2014; **14**: 97–102.

224 Robinson DG, Woerner MG, Alvir JM, *et al.* Predictors of treatment response from a first episode of schizophrenia or schizoaffective disorder. *Am J Psychiatry* 1999; **156**: 544–9.

225 Boeing L, Murray V, Pelosi A, McCabe R, Blackwood D, Wrate R. Adolescent-onset psychosis: prevalence, needs and service provision. *Br J Psychiatry* 2007; **190**: 18–26.

226 Daniel DG. Issues in selection of instruments to measure negative symptoms. *Schizophr Res* 2013; **150**: 343–5.

227 Marder SR, Davis JM, Chouinard G. The effects of risperidone on the five dimensions of schizophrenia derived by factor analysis: combined results of the North American trials. *J Clin Psychiatry* 1997; **58**: 538–46.

228 Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophr Bull* 1987; **13**: 261–76.

229 Sarkar S, Hillner K, Velligan DI. Conceptualization and treatment of negative symptoms in schizophrenia. *World J Psychiatry* 2015; **5**: 352–61.

230 Perivoliotis D, Morrison AP, Grant PM, French P, Beck AT. Negative performance beliefs and negative symptoms in individuals at ultra-high risk of psychosis: a preliminary study. *Psychopathology* 2009; **42**: 375–9.

231 Millan MJ, Fone K, Steckler T, Horan WP. Negative symptoms of schizophrenia: Clinical characteristics, pathophysiological substrates, experimental models and prospects for improved treatment. *Eur Neuropsychopharmacol* 2014; **24**: 645–92.

232 Howes OD, Kapur S. A neurobiological hypothesis for the classification of schizophrenia: type A (hyperdopaminergic) and type B (normodopaminergic). *Br J Psychiatry* 2014; **205**: 1–3.

233 NHS Digital. Hospital Episode Statistics. 2017. http://content.digital.nhs.uk/hes (accessed April 4, 2017).

234 Department for Education. National pupil database. https://www.gov.uk/government/collections/national-pupil-database (accessed Sept 3, 2017).

235 Office for National Statistics. https://www.ons.gov.uk/census/2011census (accessed Oct 18, 2016).

236 Department for Education. Schools, pupils and their characteristics: January 2014. https://www.gov.uk/government/statistics/schools-pupils-and-their-characteristics-january-2014 (accessed Oct 18, 2016).

237 Marmot M, Friel S, Bell R, Houweling TA, Taylor S. Closing the gap in a generation: health equity through action on the social determinants of health. *Lancet* 2008; **372**: 1661–9.

238 Pomerantz K, Hughes M, Thompson D. How to Reach 'Hard to Reach' Children: Improving Access, Participation and Outcomes. John Wiley & Sons, 2007.

239 Wolfe I, Lemer C, Cass H. Integrated care: a solution for improving children's health? *Arch Dis Child* 2016; **101**:992-997.

240 Kearney CA. School absenteeism and school refusal behavior in youth: A contemporary review. *Clin Psychol Rev* 2008; **28**: 451–71.

241 Whear R, Marlow R, Boddy K, *et al.* Psychiatric disorder or impairing psychology in children who have been excluded from school: A systematic review. *Sch Psychol Int* 2014; **35**: 530–43.

242 Goldman-Mellor SJ, Caspi A, Harrington H, et al. Suicide attempt in young people: A signal for long-term health care and social needs. *JAMA Psychiatry* 2014; **71**: 119–27.

243 Rodway C, Tham S-G, Ibrahim S, *et al.* Suicide in children and young people in England: a consecutive case series. *Lancet Psychiatry* 2016; **3**: 751–9.

244 NHS England. Future in mind - promoting, protecting and improving our children and young people s mental health and wellbeing. Department of Health, 2015 https:// www.gov.uk/government/uploads/system/uploads/attachment_data/ file/414024/Childrens_Mental_Health.pdf. (accessed Sept 9, 2017)

245 Clements C, Turnbull P, Hawton K, *et al.* Rates of self-harm presenting to general hospitals: a comparison of data from the Multicentre Study of Self-Harm in England and Hospital Episode Statistics. *BMJ Open* 2016; **6**: e009749.

246 Hawton K, Saunders KE, O'Connor RC. Self-harm and suicide in adolescents. *The Lancet* 2012; **379**: 2373–82.

247 Granboulan V, Roudot-Thoraval F, Lemerle S, Alvin P. Predictive factors of post-discharge follow-up care among adolescent suicide attempters. *Acta Psychiatr Scand* 2001; **104**: 31–6.

248 Taylor EA, Stansfeld SA. Children who poison themselves. II. Prediction of attendance for treatment. *Br J Psychiatry* 1984; **145**: 132–5.

249 Suominen K, Isometsä E, Martunnen M, Ostamo A, Lönnqvist J. Health care contacts before and after attempted suicide among adolescent and young adult versus older suicide attempters. *Psychol Med* 2004; **34**: 313–21.

250 Ougrin D, Zundel T, Ng A, Banarsee R, Bottle A, Taylor E. Trial of Therapeutic Assessment in London: randomised controlled trial of Therapeutic Assessment versus standard psychosocial assessment in adolescents presenting with self-harm. *Arch Dis Child* 2011; **96**: 148–53.

251 Polling C, Tulloch A, Banerjee S, *et al.* Using routine clinical and administrative data to produce a dataset of attendances at Emergency Departments following self-harm. *BMC Emerg Med* 2015; **15**: 15.

252 Woodman J, Lewis H, Cheung R, Gilbert R, Wijlaars LP. Integrating primary and secondary care for children and young people: sharing practice. *Arch Dis Child* 2015; **101**: 792–797.

253 Overy C, Reynolds LA, Tansey EM. History of the Avon Longitudinal Study of Parents and Children, C 1980-2000. Queen Mary, University of London, 2012 http://qmro.qmul.ac.uk/xmlui/handle/123456789/2827 (accessed Sept 5, 2017).

254 Bonevski B, Randell M, Paul C, *et al.* Reaching the hard-to-reach: a systematic review of strategies for improving health and medical research with socially disadvantaged groups. *BMC Med Res Methodol* 2014; **14**: 42.

255 Department for Education. National pupil database: user guide and supporting information. https://www.gov.uk/government/publications/national-pupil-database-user-guide-and-supporting-information (accessed April 4, 2017).

256 NHS Digital. The National Child Measurement Programme (NCMP). 2017; published online April 4. http://content.digital.nhs.uk/ncmp (accessed April 21, 2017).

257 EMIS Health. EMIS. http://www.emishealth.com/products/emis-web/ (accessed April 22, 2016).

258 Herrett E, Gallagher AM, Bhaskaran K, *et al.* Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* 2015; **44**: 827–36.

259 Home Office. Police National Computer (PNC). UK Government, 2017 https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/488515/PNC_v5.0_EXT_clean.pdf (accessed April 4, 2017).

260 Education and Skills Funding Agency. Individualised Learner Record (ILR). https://www.gov.uk/government/collections/individualised-learner-record-ilr (accessed April 4, 2017).

261 Public Health Research Data Forum. Enabling Data Linkage to Maximise the Value of Public Health Research Data: full report. Wellcome Trust, London. 2015 http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtp059017.pdf (accessed Nov 19, 2015).

262 Bohensky MA, Jolley D, Sundararajan V, *et al.* Data Linkage: A powerful research tool with potential problems. *BMC Health Serv Res* 2010; **10**: 346.

263 Baghal TA. Obtaining data linkage consent for children: factors influencing outcomes and potential biases. *Int J Soc Res Methodol* 2016; **19**: 623–43.

264 Health Research Authority. Section 251 and the Confidentiality Advisory Group (CAG). Health Res. Auth. http://www.hra.nhs.uk/about-the-hra/our-committees/section-251/ (accessed Oct 26, 2016).

265 Gilbert R, Lafferty R, Hagger-Johnson G, *et al.* GUILD: GUidance for Information about Linking Data sets. *J Public Health*; : 1–8.

266 Bohensky M. Bias in data linkage studies. In: Harron K, Goldstein H, Dibben C, eds. Methodological Developments in Data Linkage. John Wiley & Sons, Ltd, 2015: 63–82.

267 Lariscy JT. Differential Record Linkage by Hispanic Ethnicity and Age in Linked Mortality Studies Implications for the Epidemiologic Paradox. *J Aging Health* 2011; **23**: 1263–84.

268 Höfler M, Pfister H, Lieb R, Wittchen H-U. The use of weights to account for non-response and drop-out. *Soc Psychiatry Psychiatr Epidemiol* 2005; **40**: 291–9.

269 Little RJ, Vartivarian S. On weighting the rates in non-response weights. *Stat Med* 2003; **22**: 1589–99.

270 Downs J, Gilbert R, Hayes R, Hotopf M, Ford T. Linking health and education data to plan and evaluate services for children. *Arch Dis Child* in press. doi:10.1136/archdischild-2016-311656.

271 South London and Maudsley NHS Foundation Trust. Child and Adolescent Mental Health Services,. http://www.slam.nhs.uk/about-us/clinical-academic-groups/child-and-adolescent (accessed March 17, 2017).

272 The Information Governance Review. Information: To share or not to share? Department of Health, UK Government, 2013 https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/192572/2900774_InfoGovernance_accv2.pdf (accessed May 7, 2017).

273 HMG Cabinet office. Security policy framework - 2013 https://www.gov.uk/government/publications/security-policy-framework (accessed April 8, 2017).

274 Health Research Authority. CAG Advice and HRA/SofS Approval Decisions. Health Res. Auth. http://www.hra.nhs.uk/about-the-hra/our-committees/section-251/cag-advice-and-approval-decisions/ (accessed Sept 26, 2017).

275 Health Research Authority. Principles of Advice: Exploring the concepts of Public Interest and Reasonably Practicable. http://www.hra.nhs.uk/documents/2015/03/195369.pdf. (accessed Sept 26, 2017)

276 Information Commissioner's Office. The Guide to Data Protection. 2016 https://ico.org.uk/media/for-organisations/guide-to-data-protection-2-3.pdf (accessed Sept 26, 2017)

277 NHS Digital. https://digital.nhs.uk/home (accessed Sept 26, 2017)

278 Department for Health. NHS Information Governance Toolkit. https://www.igt.hscic.gov.uk/ (accessed April 8, 2016).

279 Hagger-Johnson G, Harron K, Goldstein H, Aldridge R, Gilbert R. Probabilistic linkage to enhance deterministic algorithms and reduce data linkage errors in hospital administrative data. *J Innov Health Inform* 2017; **24**: 234–46.

280 Shaw D. Care.data, consent, and confidentiality. *Lancet* 2014; **383**: 1205.

281 Fears R, Brand H, Frackowiak R, Pastoret P-P, Souhami R, Thompson B. Data protection regulation and the promotion of health research: getting the balance right. *QJM* 2013; **107;** 3-5

282 Hurley RF. The Decision to Trust: How Leaders Create High-Trust Organizations.: John Wiley & Sons. San Francisco, CA 2012.

283 Information Services Division Scotland. Analysis of responses to the Technical Consultation on the design of Data Sharing and Linking Service. NHS National Services Scotland, 2013 http://www.isdscotland.org/Products-and-Services/EDRIS/DSLS-consultation/DSLS-Consultation-Analysis-Report.pdf (accessed Sept 8, 2017).

284 Harron K, Hagger-Johnson G, Gilbert R, Goldstein H. Utilising identifier error variation in linkage of large administrative data sources. *BMC Med Res Methodol* 2017; **17**: 23.

285 Lundström S, Reichenberg A, Anckarsäter H, Lichtenstein P, Gillberg C. Autism phenotype versus registered diagnosis in Swedish children: prevalence trends over 10 years in general population samples. *BMJ* 2015; **350**: h1961.

286 Lai M-C, Lombardo MV, Baron-Cohen S. Autism. *Lancet* 2014; **383**: 896–910.

287 Hansen SN, Schendel DE, Parner ET. Explaining the Increase in the Prevalence of Autism Spectrum Disorders: The Proportion Attributable to Changes in Reporting Practices. *JAMA Pediatr* 2015; **169**: 56–62.

288 Camarata S. Early identification and early intervention in autism spectrum disorders: Accurate and effective? *Int J Speech Lang Pathol* 2014; **16**: 1–10.

289 Rotholz DA, Kinsman AM, Lacy KK, Charles J. Improving Early Identification and Intervention for Children at Risk for Autism Spectrum Disorder. *Pediatrics* 2017; **139**: e20161061.

290 Oono IP, Honey EJ, McConachie H. Parent-mediated early intervention for young children with autism spectrum disorders (ASD). *Cochrane Database Syst Rev* 2013; **4** : CD009774 doi: 10.1002/14651858.CD009774.pub2.

291 Hirvikoski T, Mittendorfer-Rutz E, Boman M, Larsson H, Lichtenstein P, Bölte S. Premature mortality in autism spectrum disorder. *Br J Psychiatry* 2016; **208**: 232–8.

292 Woolfenden S, Sarkozy V, Ridley G, Coory M, Williams K. A systematic review of two outcomes in autism spectrum disorder - epilepsy and mortality. *Dev Med Child Neurol* 2012; **54**: 306–12.

293 Roux AM, Shattuck PT, Cooper BP, Anderson KA, Wagner M, Narendorf SC. Postsecondary employment experiences among young adults with an autism spectrum disorder. *J Am Acad Child Adolesc Psychiatry* 2013; **52**: 931–9.

294 Shattuck PT, Narendorf SC, Cooper B, Sterzing PR, Wagner M, Taylor JL. Postsecondary Education and Employment Among Youth With an Autism Spectrum Disorder. *Pediatrics* 2012; **129**: 1042-1049

295 Magiati I, Tay XW, Howlin P. Cognitive, language, social and behavioural outcomes in adults with autism spectrum disorders: A systematic review of longitudinal follow-up studies in adulthood. *Clin Psychol Rev* 2014; **34**: 73–86.

296 Knapp M, Romeo R, Beecham J. Economic cost of autism in the UK. *Autism Int J Res Pract* 2009; **13**: 317–36.

297 Cadman T, Eklund H, Howley D, *et al.* Caregiver Burden as People With Autism Spectrum Disorder and Attention-Deficit/Hyperactivity Disorder Transition into Adolescence and Adulthood in the United Kingdom. *J Am Acad Child Adolesc Psychiatry* 2012; **51**: 879–88.

298 Pickles A, Couteur AL, Leadbitter K, *et al.* Parent-mediated social communication therapy for young children with autism (PACT): long-term follow-up of a randomised controlled trial. *The Lancet* 2016; **388**: 2501–9.

299 Simonoff E, Jones CRG, Baird G, Pickles A, Happé F, Charman T. The persistence and stability of psychiatric problems in adolescents with autism spectrum disorders. *J Child Psychol Psychiatry* 2012; **54**: 186–94.

300 Lever AG, Geurts HM. Psychiatric Co-occurring Symptoms and Disorders in Young, Middle-Aged, and Older Adults with Autism Spectrum Disorder. *J Autism Dev Disord* 2016; **46**: 1916–30.

301 Mogensen L, Mason J. The meaning of a label for teenagers negotiating identity: experiences with autism spectrum disorder. *Sociol Health Illn* 2015; **37**: 255–69.

302 Wood JJ, Drahota A, Sze K, Har K, Chiu A, Langer DA. Cognitive behavioral therapy for anxiety in children with autism spectrum disorders: a randomized, controlled trial. *J Child Psychol Psychiatry* 2009; **50**: 224–34.

303 Attwood T. Cognitive Behaviour Therapy for Children and Adults with Asperger's Syndrome. *Behav Change* 2004; **21**: 147–61.

304 Hawton K, Zahl D, Weatherall R. Suicide following deliberate self-harm: long-term follow-up of patients who presented to a general hospital. *Br J Psychiatry* 2003; **182**: 537–42.

305 Hawton K, Harriss L. Deliberate self-harm in young people: characteristics and subsequent mortality in a 20-year cohort of patients presenting to hospital. *J Clin Psychiatry* 2007; **68**: 1574–83.

306 National Collaborating Centre for Mental Health (UK). Self-Harm: Longer-Term Management. British Psychological Society. Leicester (UK): 2012 http://www.ncbi.nlm.nih.gov/books/NBK126777/ (accessed Sept 22, 2017).

307 Evans E, Hawton K, Rodham K, Psychol C, Deeks J. The Prevalence of Suicidal Phenomena in Adolescents: A Systematic Review of Population-Based Studies. *Suicide Life Threat Behav* 2005; **35**: 239–50.

308 Ribeiro JD, Franklin JC, Fox KR, *et al.* Self-injurious thoughts and behaviors as risk factors for future suicide ideation, attempts, and death: a meta-analysis of longitudinal studies. *Psychol Med* 2016; **46**: 225–36.

309 Castellví P, Lucas-Romero E, Miranda-Mendizábal A, *et al.* Longitudinal association between self-injurious thoughts and behaviors and suicidal behavior in adolescents and young adults: A systematic review with meta-analysis. *J Affect Disord* 2017; **215**: 37–48.

310 Hawton K, Rodham K, Evans E, Harriss L. Adolescents Who Self Harm: A Comparison of Those Who Go to Hospital and Those Who Do Not. *Child Adolesc Ment Health* 2009; **14**: 24–30.

311 Olfson M, Gameroff MJ, Marcus SC, Greenberg T, Shaffer D. Emergency treatment of young people following deliberate self-harm. *Arch Gen Psychiatry* 2005; **62**: 1122–8.

312 Owens C, Hansford L, Sharkey S, Ford T. Needs and fears of young people presenting at accident and emergency department following an act of self-harm: secondary analysis of qualitative data. *Br J Psychiatry* 2016; **208**: 286–91.

313 Gairin I, House A, Owens D. Attendance at the accident and emergency department in the year before suicide: retrospective study. *Br J Psychiatry* 2003; **183**: 28–33.

314 Hawton PK, Bergen H, Casey D, *et al.* Self-harm in England: a tale of three cities. *Soc Psychiatry Psychiatr Epidemiol* 2007; **42**: 513–21.

315 Hawton K, Rodham K, Evans E, Weatherall R. Deliberate self harm in adolescents: self report survey in schools in England. *BMJ* 2002; **325**: 1207–11.

316 Moran P, Coffey C, Romaniuk H, *et al.* The natural history of self-harm from adolescence to young adulthood: a population-based cohort study. *Lancet* 2012; **379**: 236–43.

317 Brunner R, Kaess M, Parzer P, *et al.* Life-time prevalence and psychosocial correlates of adolescent direct self-injurious behavior: A comparative study of findings in 11 European countries. *J Child Psychol Psychiatry* 2014; **55**: 337–48.

318 Fisher HL, Moffitt TE, Houts RM, Belsky DW, Arseneault L, Caspi A. Bullying victimisation and risk of self harm in early adolescence: longitudinal cohort study. *BMJ* 2012; **344**: e2683.

319 Rhodes AE, Boyle MH, Bethell J, *et al.* Child maltreatment and onset of emergency department presentations for suicide-related behaviors. *Child Abuse Negl* 2012; **36**: 542–51.

320 Kretschmar JM, Flannery DJ. Displacement and Suicide Risk for Juvenile Justice-Involved Youth with Mental Health Issues. *J Clin Child Adolesc Psychol* 2011; **40**: 797–806.

321 Evans R, Hurrell C. The role of schools in children and young people's self-harm and suicide: systematic review and meta-ethnography of qualitative research. *BMC Public Health* 2016; **16**. doi:10.1186/s12889-016-3065-2.

322 Cassidy S, Bradley P, Robinson J, Allison C, McHugh M, Baron-Cohen S. Suicidal ideation and suicide plans or attempts in adults with Asperger's syndrome attending a specialist diagnostic clinic: a clinical cohort study. *Lancet Psychiatry* 2014; **1**: 142–7.

323 Croen LA, Zerbo O, Qian Y, *et al.* The health status of adults on the autism spectrum. *Autism* 2015; **19**: 814–23.

324 Tantam D, Girgis S. Recognition and treatment of Asperger syndrome in the community. *Br Med Bull* 2009; **89**: 41–62.

325 Fee V, Matson. Self-injurious behavior: Analysis, assessment, and treatment. In: Definition, classification, and taxonomy n J. K. Luiselli, J. L. Matson, & N. Singh (Eds.) Springer, New York. 1992: 3–20.

326 Soke GN, Rosenberg SA, Hamman RF, *et al.* Factors Associated with Self-Injurious Behaviors in Children with Autism Spectrum Disorder: Findings from Two Large National Samples. *J Autism Dev Disord* 2017; **47**: 285–96.

327 Schroeder SR, Marquis JG, Reese RM, *et al.* Risk factors for self-injury, aggression, and stereotyped behavior among young children at risk for intellectual and developmental disabilities. *Am J Intellect Dev Disabil* 2014; **119**: 351–70.

328 Myers SM, Johnson CP. Management of Children With Autism Spectrum Disorders. *Pediatrics* 2007; **120**: 1162–82.

329 Mars B, Heron J, Crane C, *et al.* Differences in risk factors for self-harm with and without suicidal intent: Findings from the ALSPAC cohort. *J Affect Disord* 2014; **168**: 407–14.

330 Storch EA, Sulkowski ML, Nadeau J, *et al.* The phenomenology and clinical correlates of suicidal thoughts and behaviors in youth with autism spectrum disorders. *J Autism Dev Disord* 2013; **43**: 2450–9.

331 Mayes SD, Gorman AA, Hillwig-Garcia J, Syed E. Suicide ideation and attempts in children with autism. *Res Autism Spectr Disord* 2013; **7**: 109–19.

332 NICE. Self-harm in over 8s: short-term management and prevention of recurrence (Clincial Guidance 16). National Institute for Health and Care Excellence. London, UK. 2004 https://www.nice.org.uk/guidance/cg16 (accessed Sept 8, 2017).

333 Nock MK, Borges G, Bromet EJ, Cha CB, Kessler RC, Lee S. Suicide and Suicidal Behavior. *Epidemiol Rev* 2008; **30**: 133–54.

334 Department for Education. Children with Special Education Needs 2013: An Analysis. Department for Education, 2013 https://www.gov.uk/government/collections/statisticsspecial- educational-needs-sen. (accessed Sept 8, 2017)

335 Baron-Cohen S, Scott FJ, Allison C, *et al.* Prevalence of autism-spectrum conditions: UK school-based population study. *Br J Psychiatry* 2009; **194**: 500–9.

336 Bresin K, Schoenleber M. Gender differences in the prevalence of nonsuicidal self-injury: A meta-analysis. *Clin Psychol Rev* 2015; **38**: 55–64.

337 StataCorp. Stata user's guide release 14. Texas: Stata Press Publication, 2015 http://www.stata.com/manuals14/u.pdf.

338 Sterne JAC, White IR, Carlin JB, *et al.* Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009; **338**: b2393.

339 White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med* 2011; **30**: 377–99.

340 Department for Education. Apply for free school meals. https://www.gov.uk/apply-free-school-meals (accessed Aug 16, 2017).

341 Ford T, Hamilton H, Meltzer H, Goodman R. Predictors of Service Use for Mental Health Problems Among British Schoolchildren. *Child Adolesc Ment Health* 2008; **13**: 32–40.

342 Mandy W, Lai M-C. Annual Research Review: The role of the environment in the developmental psychopathology of autism spectrum condition. *J Child Psychol Psychiatry* 2016; **57**: 271–92.

343 Mandy W, Murin M, Baykaner O, *et al.* The transition from primary to secondary school in mainstream education for children with autism spectrum disorder. *Autism* 2016; **20**: 5–13.

344 Picci G, Scherf KS. A. Two-Hit Model of Autism: Adolescence as the Second Hit. *Clin Psychol Sci* 2015; **3**: 349–71.

345 Kerns CM, Kendall PC, Zickgraf H, Franklin ME, Miller J, Herrington J. Not to Be Overshadowed or Overlooked: Functional Impairments Associated With Comorbid Anxiety Disorders in Youth With ASD. *Behav Ther* 2015; **46**: 29–39.

346 Frazier JA, Doyle R, Chiu S, Coyle JT. Treating a Child With Asperger's Disorder and Comorbid Bipolar Disorder. *Am J Psychiatry* 2002; **159**: 13–21.

347 Frazier JA, Biederman J, Bellordre CA, *et al.* Should the diagnosis of Attention-Deficit/ Hyperactivity disorder be considered in children with Pervasive Developmental Disorder? *J Atten Disord* 2001; **4**: 203–11.

348 Patel V, Flisher AJ, Hetrick S, McGorry P. Mental health of young people: a global public-health challenge. *Lancet* 2007; **369**: 1302–13.

349 Chiri G, Warfield ME. Unmet Need and Problems Accessing Core Health Care Services for Children with Autism Spectrum Disorder. *Matern Child Health J* 2012; **16**: 1081–91.

350 Autistica-JL alliance. Your questions shaping the future of autism research: an Autistica-James Lind Alliance Research Priority Setting Partnership. Autistica. London. 2017 http://www.jla.nihr.ac.uk/priority-setting-partnerships/autism/ (accessed July 8, 2017).

351 Claire Parker, Ruth Marlow, Marc Kastner, *et al.* The 'Supporting Kids, Avoiding Problems' (SKIP) study: relationships between school exclusion, psychopathology, development and attainment – a case control study. *J Child Serv* 2016; **11**: 91–110.

352 Allely CS. The association of ADHD symptoms to self-harm behaviours: a systematic PRISMA review. *BMC Psychiatry* 2014; **14**: 133.

353 Burrows S, Laflamme L. Socioeconomic disparities and attempted suicide: state of knowledge and implications for research and prevention. *Int J Inj Contr Saf Promot* 2010; **17**: 23–40.

354 Jablonska B, Lindberg L, Lindblad F, Hjern A. Ethnicity, socio-economic status and self-harm in Swedish youth: a national cohort study. *Psychol Med* 2009; **39**: 87–94.

355 Bentley KH, Cassiello-Robbins CF, Vittorio L, Sauer-Zavala S, Barlow DH. The association between nonsuicidal self-injury and the emotional disorders: A meta-analytic review. *Clin Psychol Rev* 2015; **37**: 72–88.

356 Hawton K, Hall S, Simkin S, *et al.* Deliberate self-harm in adolescents: a study of characteristics and trends in Oxford, 1990-2000. *J Child Psychol Psychiatry* 2003; **44**: 1191–8.

357 Diggins E, Kelley R, Cottrell D, House A, Owens D. Age-related differences in self-harm presentations and subsequent management of adolescents and young adults at the emergency department. *J Affect Disord* 2017; **208**: 399–405.

358 Madge N, Hewitt A, Hawton K, *et al.* Deliberate self-harm within an international community sample of young people: comparative findings from the Child & Adolescent Self-harm in Europe (CASE) Study. *J Child Psychol Psychiatry* 2008; **49**: 667–77.

359 Public Health England. Public Health Profiles : self harm admission rates per 100,00 (Age 10-24). https://fingertips.phe.org.uk/ (accessed Aug 16, 2017).

360 Public Health England. Child and maternal health data and intelligence: a guide for health professionals. https://www.gov.uk/guidance/child-and-maternal-health-data-and-intelligence-a-guide-for-health-professionals (accessed Aug 9, 2017).

361 Majid M, Tadros M, Tadros G, Singh S, Broome MR, Upthegrove R. Young people who self-harm: a prospective 1-year follow-up study. *Soc Psychiatry Psychiatr Epidemiol* 2015; **51**: 171–81.

362 Achenbach TM, Ivanova MY, Rescorla LA, Turner LV, Althoff RR. Internalizing/Externalizing Problems: Review and Recommendations for Clinical and Research Applications. *J Am Acad Child Adolesc Psychiatry* 2016; **55**: 647–56.

363 Masten AS, Roisman GI, Long JD, *et al.* Developmental cascades: linking academic achievement and externalizing and internalizing symptoms over 20 years. *Dev Psychol* 2005; **41**: 733–46.

364 Moilanen KL, Shaw DS, Maxwell KL. Developmental cascades: externalizing, internalizing, and academic competence from middle childhood to early adolescence. *Dev Psychopathol* 2010; **22**: 635–53.

365 Patalay P, Deighton J, Fonagy P, Wolpert M. The relationship between internalising symptom development and academic attainment in early adolescence. *PloS One* 2015; **10**: e0116821.

366 Downs J, Velupillai S, Gkotsis G, *et al.* Detection of Suicidality in Adolescents with Autism Spectrum Disorders: Developing a Natural Language Processing Approach for Use in Electronic Health Records. *Proc Am Med Inform Assoc* 2017.

367 Hodkinson A. Key Issues in Special Educational Needs and Inclusion. SAGE, 2015.

368 Howes OD, McCutcheon R, Agid O, *et al.* Treatment-Resistant Schizophrenia: Treatment Response and Resistance in Psychosis (TRRIP) Working Group Consensus Guidelines on Diagnosis and Terminology. *Am J Psychiatry* 2016; **174**: 216–29.

369 Victora C. What's the denominator? *Lancet* 1993; **342**: 97–9.

370 Larsen MD, Cars T, Hallas J. A MiniReview of the Use of Hospital-based Databases in Observational Inpatient Studies of Drugs. *Basic Clin Pharmacol Toxicol* 2013; **112**: 13–8.

371 Kaplan RM, Chambers DA, Glasgow RE. Big data and large sample size: a cautionary note on the potential for bias. *Clin Transl Sci* 2014; **7**: 342–6.

372 Stanley F, Glauert R, McKenzie A, O'Donnell M. Can Joined-Up Data Lead to Joined-Up Thinking? The Western Australian Developmental Pathways Project. *Healthc Policy* 2011; **6**: 63–73.

373 Commission on the Future of Health and Social Care in England. The UK private health market. The King's Fund, 2014 https://www.kingsfund.org.uk/sites/default/files/media/commission-appendix-uk-private-health-market.pdf (accessed Jan 8, 2017).

374 Mok PLH, Webb RT, Appleby L, Pedersen CB. Full spectrum of mental disorders linked with childhood residential mobility. *J Psychiatr Res* 2016; **78**: 57–64.

375 Westreich D. Berkson's bias, selection bias, and missing data. *Epidemiol Camb Mass* 2012; **23**: 159–64.

376 Campbell F, Biggs K, Aldiss SK, *et al.* Transition of care for adolescents from paediatric services to adult health services. In: Cochrane Database of Systematic Reviews. John Wiley & Sons, Ltd, 2016. http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD009794.pub2/abstract (accessed Sept 11, 2017).

377 Green J. Annotation: The therapeutic alliance – a significant but neglected variable in child mental health treatment studies. *J Child Psychol Psychiatry* 2006; **47**: 425–35.

378 Cirulli G. Clozapine prescribing in adolescent psychiatry: Survey of prescribing practice in in-patients units. *Psychiatr Bull* 2005; **29**: 377–80.

379 Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychol Bull* 1955; **52**: 281–302.

380 Elbe D, Mc Glanaghy E, Oberlander TF. Chapter 3 - Do We Know If They Work and If They Are Safe: Second-Generation Antipsychotics for Treatment of Autism Spectrum Disorders and Disruptive Behavior Disorders in Children and Adolescents. In: Pietro ND, Illes J, (eds.) The Science and Ethics of Antipsychotic Use in Children. Academic Press. San Diego: 2015: 27–64.

381 Ipsos MORI. Dialogue on Data. Ipsos MORI Social Research Institute, 2014 http://www.ipsos-mori.com/DownloadPublication/1652_sri-dialogue-on-data-2014.pdf.

382 Velupillai S, Suominen H, Liakata M, *et al.* The Interplay of Evaluating Natural Language Processing Approaches and Clinical Outcomes Research. *J Biomed Inform* 2017; **under review**.

383 UK CRIS C. CRIS Network Members. CRIS Netw. https://crisnetwork.co/members (accessed Sept 14, 2017).

384 Robertson A, Cresswell K, Takian A, *et al.* Implementation and adoption of nationwide electronic health records in secondary care in England: qualitative analysis of interim results from a prospective national evaluation. *BMJ* 2010; **341**: c4564.

385 National Advisory Group on Health Information Technology in England. Using information technology to improve the NHS. Department of Health. London. 2016 https://www.gov.uk/government/publications/using-information-technology-to-improve-the-nhs(accessed Sept 14, 2017)..

386 Evans R. The woman falsely labelled alcoholic by the NHS. The Guardian. 2006; published online Nov 2. http://www.theguardian.com/society/2006/nov/02/health.epublic (accessed Sept 15, 2017).

387 McGrath-Lone L, Woodman J, Gilbert R. Safeguarding children and improving their care in the UK. *The Lancet* 2015; **386**: 1630.

388 Muir BM. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics* 1994; **37**: 1905–22.

APPENDIX A

Cohort table describing the year of study entry and mean duration under follow for 113,543 secondary school age pupils in South London

Year of study entry	AGE and follow-up characteristics													
	11		12		13		14		15		16		17	
	n	Mean FU	n	Mean FU	n	Mean FU	n	Mean FU	n	Mean FU	n	Mean FU	n	Mean FU
2009	20708	3.83	11351	3.99	11199	3.99	10978	3.51	10348	2.52	7672	1.6	3775	0.53
2010	11814	2.75												
2011	11441	1.74												
2012	11301	0.75												
2013	2948	0.12												

*FU : duration of follow up in years

**Much larger cohort as numbers reflect all children who turned 11 between 1st January 2009 and 31st December 2009, as well those who were already aged 11. (Follow-up period started for this group at the date of 11th birthday)

APPENDIX B

Health Research Authority
NRES Committee South Central - Oxford C
Bristol REC Centre
Level 3, Block B
Whitefriars Building
Lewins Mead
Bristol
BS1 2NT

Telephone: 01173421392

21 May 2013

Dear Prof. Stewart

Title of the Database:	**South London and Maudsley Biomedical Research Centre Clinical Case Register**
REC reference:	**08/H0606/71**
Amendment number:	**Substantial Amendment 2: Link data from CRIS to information contained within the NPD database.**
Amendment date:	**09 April 2013**
IRAS project ID:	

The above amendment was reviewed on 13 May 2013 by the Sub-Committee in correspondence.

Ethical opinion

The members of the Committee taking part in the review gave a favourable ethical opinion of the amendment on the basis described in the notice of amendment form and supporting documentation.

Approved documents

The documents reviewed and approved at the meeting were:

Document	Version	Date
Letter confirming Past Approval of System Level Security Policy by NIGB ECC		11 April 2011
Section 251 Application to be Submitted		
Biomedica; Research Centre Data Linkage Service System Level Security Policy		21 February 2011
Research Protocol to be Submitted to NIGB ECC		09 April 2013
Notice of Substantial Amendment	Substantial Amendment 2:	09 April 2013
Covering Letter		09 April 2013
Calidicott Guardian Letter of Approval		11 February 2013
Protocol for Management of the Database		02 May 2013

Membership of the Committee

The members of the Ethics Committee who took part in the review are listed on the attached sheet.

Statement of compliance

The Committee is constituted in accordance with the Governance Arrangements for Research Ethics Committees and complies fully with the Standard Operating Procedures for Research Ethics Committees in the UK.

We are pleased to welcome researchers and R & D staff at our NRES committee members' training days – see details at http://www.hra.nhs.uk/hra-training/

08/H0606/71	**Please quote this number on all correspondence**

Yours sincerely

P. P.

Professor Nigel Wellman
Chair

E-mail: nrescommittee.southcentral-oxfordc@nhs.net

Enclosures: *List of names and professions of members who took part in the review*

NRES Committee South Central - Oxford C

Attendance at Sub-Committee of the REC meeting on 13 May 2013

Committee Members:

Name	*Profession*	*Present*	*Notes*
Dr Avinash Gupta	Clinical Research Fellow	Yes	
Professor Nigel Wellman	Professor of Health and Human Sciences	Yes	

Also in attendance:

Name	*Position (or reason for attending)*
Miss Lidia Gonzalez	Administration Assistant

NHS
Health Research Authority

Confidentiality Advisory Group

Professor Matthew Hotopf
Institute of Psychiatry
Kings College London
Weston Education Centre
Cutcombe Road
SE5 9RJ

Skipton House
80 London Road
London
SE1 6LH

Tel: 020 797 22557
Email: HRA.CAG@nhs.net

10 March 2014

Dear Professor Hotopf

Study title: **SLAM CAMHS CRIS linkage with DfE National Pupil Database**
CAG reference: **CAG 9-08(a)/2014**
REC number: **08/H0606/71**

Thank you for your research application, submitted for approval under Regulation 5 of the Health Service (Control of Patient Information) Regulations 2002 to process patient identifiable information without consent. Approved applications enable the data controller to provide specified information to the applicant for the purposes of the relevant activity, without being in breach of the common law duty of confidentiality, although other relevant legislative provisions will still be applicable.

The role of the Confidentiality Advisory Group (CAG) is to review applications submitted under these Regulations and to provide advice to the Health Research Authority on whether an application should be approved, and if so, any relevant conditions. This application was considered on 09 January 2014.

Health Research Authority approval decision

The Health Research Authority, having considered the advice from the Confidentiality Advisory Group as set out below, has determined the following:

1. The application is approved, subject to compliance with the standard and specific conditions of approval.

This letter should be read in conjunction with the outcome letter dated 23 January 2014.

Context

Purpose of application

This application from Kings College London set out the purpose of linking and anonymising child and adolescent mental health (CAMHS) clinical data from the South London and Maudsley NHS Foundation Trust (SLAM) Biomedical Research Centre (BRC) Case Register Interactive Search (CRIS) system and educational performance data from Department of Education (DfE) National Pupil Database (NPD).

All children aged between 5 and 17 who were referred to CAMHS services between January 2008 and December 2013, (approx 35,000) would be included in order to aid health and education policy makers by providing information on the frequency and characteristics of children referred to CAMHS.

A recommendation for class 1, 4, 5 and 6 support was requested to cover DfE access to demographic data only from CAMHS.

Confidential patient information requested

Access was requested to first name, last name, date of birth and address.

Background

This application had previously been considered by the CAG predecessor, the Ethics and Confidentiality Committee (ECC 8-04 (a)/2013) and it was advised at that time that the application could not be supported. The following issues were raised:

1. The description of purpose section within the application form should be revised to ensure that this reflects both the medical purpose and public interest in the activity taking place. This should include examples of the medical research that will be undertaken using the data.

2. Consideration should be given to whether the HSCIC's TDLS could undertake linkages.

3. The patient information leaflet should be revised to reflect the disclosure of identifiable data to other organisations.

4. Further information in relation to the governance controls within DfE, including justification for the disclosure of NHS number, should be provided.

Further information in relation to all points listed above was provided in the submission to the CAG meeting on the 9 January 2014.

Confidentiality Advisory Group advice

In line with the considerations above, the CAG agreed that the minimum criteria under the Regulations appeared to have been met, and therefore advised recommending *provisional* support to the Health Research Authority, subject to further information being submitted in line with the request for clarification and compliance with the specific and standard conditions of support as set out below.

Further information was provided by the applicant on the 10 February 2014 in response to the request for clarification and is summarised below in bold.

Request for clarification

1. Please provide further information in relation to governance arrangements within DfE, including confirmation of retention period for CAMHS data and how access will be restricted within DfE. **It was confirmed that the data will not be retained by the Department for more than 60 days. The DfE contractors directly involved in the matching of personal identifiers from SLaM (in total one DfE contracted staff), will not know the origins of the data nor have direct contact with SLaM CDLS staff.**

2. Please confirm what confidentiality agreements are in place to ensure that DfE staff remain aware of their responsibilities when processing personal data. **A copy of DfE's**

personal information charter was forwarded to the Confidentiality Advisory Group.

Specific conditions of support

1. Favourable opinion from Research Ethics Committee. **Confirmed 25 February 2014.**

2. Confirmation of suitable security arrangements via IG Toolkit submission. **Arrangements at DfE confirmed as satisfactory on 13 November 2013.**

As the above conditions have been accepted and/or met, this letter provides confirmation of final approval. I will arrange for the register of approved applications on the HRA website to be updated with this information.

Annual review

Please note that your approval is subject to submission of an annual review report to show how you have met the conditions or report plans, and action towards meeting them. It is also your responsibility to submit this report on the anniversary of your final approval and to report any changes such as to the purpose or design of the proposed activity, or to security and confidentiality arrangements. We are also streamlining the process to facilitate the service we provide to applicants. This means that annual reviews will be batched and reviewed on the last day of the preceding month before the date of approval. An annual review should therefore be provided no later than 28 February 2015 and preferably 4 weeks before this date.

Please do not hesitate to contact me if you have any queries following this letter. I would be grateful if you could quote the above reference number in all future correspondence.

Reviewed documents

The documents reviewed at the meeting were:

Document	*Version*	*Date*
Covering Letter from Dr Johnny Downs		6/12/2013
Research Ethics Committee favourable opinion letter		21/05/2013
IRAS application form		06/12/2013
Patient Information Leaflets		06/12/2013
Research protocol substantial amendment		9/04/2013
Case for Support		06/12/2013
Caldicott Guardian support letter		11/02/2013
BMC Medical Informatics and Decision Making article		11/07/2013
Department for Education support letters		17/09/2013 and 21/11/2013

Membership of the Group

The members of the Confidentiality Advisory Group who were present at the consideration of this item are listed below.

Dr Murat Soncul declared a conflicting interest in the application as previously named within the application to ECC and an employee of SLAM.

Feedback

You are invited to give your view of the service provided by the Confidentiality Advice Team and the application procedure in general by completion of this survey https://www.surveymonkey.com/s/KPRFK5T. We would be grateful if you could take some time to provide your feedback.

With the Group's best wishes for the success of this project.

Yours sincerely

Claire Edgeworth
Deputy Confidentiality Advice Manager

Email: HRA.CAG@nhs.net

Copy to: nrescommittee.southcentral-oxfordc@nhs.net

Enclosures: *List of members who were present at the meeting and those who submitted written comments*

Standard conditions of approval

Confidentiality Advisory Group
Attendance at meeting on 18 April 2013

Group members

Name	Capacity
Dr Mark Taylor (Chair)	Lay
Dr Charlotte Augst	
Dr Kambiz Boomla	
Dr Tony Calland	
Dr Robert Carr	
Mr Paul Charlton	Lay
Ms Madeleine Colvin	
Professor Julia Hippisley-Cox	
Dr Patrick Coyle	
Dr Tricia Cresswell (vice-chair)	
Mr Anthony Kane	Lay
Professor Jennifer Kurinczuk	
Ms Clare Sanderson	
Mr C. Marc Taylor	
Ms Gillian Wells	Lay
Dr Christopher Wiltsher	Lay
Mr Terence Wiseman	Lay

NHS
Health Research Authority
Confidentiality Advisory Group

Standard conditions of approval

The approval provided by the Health Research Authority is subject to the following standard conditions.

The applicant will ensure that:

1. The specified patient identifiable information is only used for the purpose(s) set out in the application.

2. Confidentiality is preserved and there are no disclosures of information in aggregate or patient level form that may inferentially identify a person, nor will any attempt be made to identify individuals, households or organisations in the data.

3. Requirements of the Statistics and Registration Services Act 2007 are adhered to regarding publication when relevant.

4. All staff with access to patient identifiable information have contractual obligations of confidentiality, enforceable through disciplinary procedures.

5. All staff with access to patient identifiable information have received appropriate ongoing training to ensure they are aware of their responsibilities.

6. Activities are consistent with the Data Protection Act 1998.

7. Audit of data processing by a designated agent is facilitated and supported.

8. The wishes of patients who have withheld or withdrawn their consent are respected.

9. The Confidentiality Advice Team is notified of any significant changes (purpose, data flows, data items, security arrangements) prior to the change occurring.

10. An annual report is provided no later than 12 months from the date of your final confirmation letter.

11. Any breaches of confidentiality / security around this particular flow of data should be reported to CAG within 10 working days, along with remedial actions taken / to be taken.

NHS
Health Research Authority

Professor Robert Stewart
Professor of Psychiatric Epidemiology & Clinical Informatics
Department of Psychological Medicine
Institute of Psychiatry, King's College London
De Crespigny Park
London SE5 8AF
United Kingdom

Skipton House
80 London Road
London
SE1 6LH

Tel: 020 797 22557
Email: HRA.CAG@nhs.net

10 August 2016

Dear Professor Stewart

Study title: **SLAM IG Clinical Dataset Linking Service**
CAG reference: **ECC 3-04(f)/2011**
REC number: **08/H0606//71**

Thank you for your amendment request to the above research application, submitted for approval under Regulation 5 of the Health Service (Control of Patient Information) Regulations 2002 to process patient identifiable information without consent. Approved applications enable the data controller to provide specified information to the applicant for the purposes of the relevant activity, without being in breach of the common law duty of confidentiality, although other relevant legislative provisions will still be applicable.

The role of the Confidentiality Advisory Group (CAG) is to review applications submitted under these Regulations and to provide advice to the Health Research Authority on whether an application should be approved, and if so, any relevant conditions.

Health Research Authority approval decision

The Health Research Authority, having considered the advice from the Confidentiality Advisory Group as set out below, has determined the following:

1. The amendment is approved, subject to compliance with the standard conditions of support.

Context

This research application from the South London & Maudsley NHS Foundation Trust set out the purpose of investigating the associations between specific mental disorders in secondary mental health care (schizophrenia, schizoaffective disorder, bipolar disorder and dementia) and physical illness. This would use a new linked dataset containing health records for patients with these disorders from the SLAM BRC Case Register Interactive Search (CRIS) and general hospital records from the English national Hospital Episode Statistics (HES) database. Review of this application was sought so as to provide a legitimate basis for the processing of this patient identifiable information; to effectively test this 'honest broker' capability and to permit the linkage and subsequent anonymisation. This required access to name, date of birth, sex, address, postcode and NHS Number.

Amendment request

This application is to use the HES data already linked to CRIS to investigate presentations to hospital due to suicidality and self-harm by young people. It extends the scope of the original application by requesting to use data relating to children under the age of 18 whereas the original application only covered adults. A previous amendment dated 28/11/2013 has extended permission to allow the use of under 18s data for a project looking at epilepsy outcomes in children with autistic spectrum disorder treated with psychotropic medication.

The planned project would use HES A&E data for patients known to SLaM to ascertain A&E attendances, with linked clinical record data being used to get information about the reason for attendance. It would also use HES inpatient data to ascertain admissions coded as due to self-harm, and physical health co-morbidities in those presenting with self-harm and suicidal behaviour. It would use HES data on all residents in the area covered by SLaM to provide a comparison population.

Confidentiality Advisory Group advice

The amendment requested was forwarded to the Chair, who determined that as the amendment involved both mental health data and the use of health care data relating to minors, it required review by Sub-Committee.

The Sub-Committee noted that the amendment involved the use of a control group, who would not directly benefit from this use of their data. However, it was agreed that there was a strong public interest in this research, which the Confidentiality Advisory Group had already supported for a different (and potentially less relevant) group of patients.

The Sub-Committee considered the patient information provided, which had been previously approved by the Confidentiality Advisory Group for children and young people. It was deemed adequate for this project.

Confidentiality Advisory Group conclusion

In line with the considerations above, the Chair agreed that the minimum criteria under the Regulations appeared to have been met for this amendment, and therefore advised recommending support to the Health Research Authority.

Specific conditions of support

1. Confirmation of suitable security arrangements via IG Toolkit submission.
2. Confirmation of a favourable opinion from a Research Ethics Committee.

Reviewed documents

Document	*Version*	*Date*
Cover letter		17 June 2016
Amendment Request form		17 June 2016
CRIS CAMHS leaflet		21 February 2014
CRIS data linkages webpage		
Supplementary information (summary of study and previous amendments)		

Please do not hesitate to contact me if you have any queries following this letter. I would be grateful if you could quote the above reference number in all future correspondence.

Yours sincerely

Confidentiality Advisor
On behalf of the Health Research Authority

Email: HRA.CAG@nhs.net

Enclosures: Standard conditions of approval

NHS
Health Research Authority

Standard conditions of approval

The approval provided by the Health Research Authority is subject to the following standard conditions.

The applicant will ensure that:

1. The specified patient identifiable information is only used for the purpose(s) set out in the application.

2. Confidentiality is preserved and there are no disclosures of information in aggregate or patient level form that may inferentially identify a person, nor will any attempt be made to identify individuals, households or organisations in the data.

3. Requirements of the Statistics and Registration Services Act 2007 are adhered to regarding publication when relevant.

4. All staff with access to patient identifiable information have contractual obligations of confidentiality, enforceable through disciplinary procedures.

5. All staff with access to patient identifiable information have received appropriate ongoing training to ensure they are aware of their responsibilities.

6. Activities are consistent with the Data Protection Act 1998.

7. Audit of data processing by a designated agent is facilitated and supported.

8. The wishes of patients who have withheld or withdrawn their consent are respected.

9. The Confidentiality Advice Team is notified of any significant changes (purpose, data flows, data items, security arrangements) prior to the change occurring.

10. An annual report is provided no later than 12 months from the date of your final confirmation letter.

11. Any breaches of confidentiality / security around this particular flow of data should be reported to CAG within 10 working days, along with remedial actions taken / to be taken.

V2.0

Clinical Data Linkage Service Memorandum of Understanding

SUMMARY

South London and Maudsley NHS foundation trust (SLaM) Information Governance (IG) Clinical Data Linkage Service (CDLS) provides an independent, safe and secure service to link and/or host and/or store clinical data from different sources for parties wanting to use linked data to perform health related research.

As a trusted third party data linkage service and safe haven CDLS operates independently from the service users, who want to use linked clinical data for research, and data controllers, who are responsible for the source clinical data. Rather, CDLS acts as an honest broker between these parties, providing researchers access to essential clinical data for health research within a framework that ensures the highest, up-to-date Information Governance (IG) standards required by Data Controllers are met.

This MoU forms an agreement between CDLS and Data Controllers of data sources being linked. It specifies CDLS's data stewardship function and defines responsibility and liability for performance of project related data activities.

This document and its schedules have a template structure which to meet specific project requirements.

The Main Body of this document details the names of contact details for Parties involved in the project. It defines the key terms used throughout the document. It stipulates the general responsibilities and liabilities of the Parties acting in the capacity of either Data Controller or Data Processor in relation to this project.

Schedule 1 is comprised of the s.251 application approved for this project. It defines the specific and detailed process agreed between the Parties in transferring, linking, hosting and accessing relevant data sets. This stipulates the rationale for the project, services required for the project, identifiers used (where appropriate), how the data will be processed and for how long it will be held or hosted.

Schedule 2 requires names and contact details for the purpose of communicating project administration.

Schedule 3 describes the relevant Standards and Policies CDLS adheres to, e.g.

- SLaM Information Governance (IG) and Information and Communication Technology (ICT) security policies.
- The Department of Health (DoH) IG Toolkit for Mental Health Trusts standards.

CDLS MEMORANDUM OF UNDERSTANDING

DATED: 31st December 2014

BETWEEN:

(1) **SOUTH LONDON AND MAUDSLEY NHS FOUNDATION TRUST** (SLaM) of Trust Head Quarters, The Maudsley Hospital, Denmark Hill, London, SE5 8AZ; and

(2) **DEPARTMENT FOR EDUCATION** (DfE) of Sanctuary Buildings, 20 Great Smith St, London SW1P 3BT

INTRODUCTION

(A) South London and Maudsley NHS Foundation Trust ("SLaM") is the Data Controller of the CRIS dataset.

(B) In addition to its role as Data Controller for CRIS, SLaM is the host organisation for an operational Data Processor unit known as the Clinical Data Linkage Service ("CDLS"). CDLS forms part of the SLaM Information Governance ("IG") and Information and Communications Technology Department ("ICT").

(C) Department for Education (DfE) is the Data Controller of the National Pupil Database (NPD).

(D) In addition to its role as Data Controller for NPD, DfE will act as Data Processor for activities described in the Process.

(E) To meet the above objectives, South London and Maudsley NHS Foundation Trust and the Department for Education (The Parties) agree to abide by the terms and conditions of this Agreement.

1. **DEFINITIONS**

1.1 The following terms have the following meanings:

"Agreement"	means this agreement including the schedules and any other documents which are expressly identified herein as applicable to this agreement;
"Authorised Representative"	means the persons authorised to represent the Parties in the performance of this Agreement, as described in Schedule 2 (or their replacements as notified in writing from time to time);
"CDLS"	means the Clinical Data Linkage Service, an operational unit at SLaM which will perform Data Processing for and on behalf of SLaM;
"CDLS Standards"	means specifically the Standards set out in Part A: of Schedule 3 to this Agreement;
"Data Activities"	means the actions to be performed by the Parties in respect of the Identifiers, as described in Clause 5.1 and pursuant to the Process;
"Data Controller"	means (subject to subsection (4) of the DPA) a person who (either alone or jointly or in common with other persons) determines the purposes for which and the manner in which any Personal Data are, or are to be, processed;
"Data Processor"	in relation to Personal Data, means any person (other than an employee of the Data Controller) who processes the data on behalf of the Data Controller;
"Data User/s"	refers to the third party entities or named individuals detailed in Part A: Section 17 of the s.251 application in Schedule 1 to this Agreement;
"DPA"	means the Data Protection Act 1998 (as amended);

"FOIA"	means the Freedom of Information Act 2000 (as amended);
"Identifiers"	means certain data provided to a Data Processor to be used in the process of linking independent data sets and which falls within the specific data categories detailed in Part A; Section 11 and 12-2 of the s.251 application in Schedule 1 to this Agreement;
"Output"	means the data extracted for Data Users by CDLS from the data described in Part A: Section 6 of the s.251 application within Schedule 1 to this Agreement, according to the process described in Section 10 and Section 18 to 25 of the same;
"Parties"	refers collectively to those mentioned in (1) and (2) of this MoU;
"Personal Data"	shall have the meaning set out in the DPA;
"Policies"	means the generally applicable Policies described in Part A and B of Schedule 3 together with any additional specific Policies which are applicable to the Data Activities;
"Process"	means the process described in Part B Section 8 of the s.251 application in Schedule 1 to this Agreement;
"Purpose"	means only the purpose described Part A: Section 7 of the s.251 application in Schedule 1 to this Agreement;
s.251	means Section 251 of the National Health Service Act 2006;
"Standards"	means the generally applicable Standards described in Part A and B of Schedule 3 together with any additional specific Standards which are applicable to the Data Activities;
"Term"	means the term of this Agreement according to Clause 18 and detailed in Schedule 1 to this agreement.

2. GOOD FAITH

2.1 The parties recognise that it is impracticable to make provision for every contingency which may arise during the life of this Agreement and they declare it to be their intention that this Agreement shall operate between the Parties with fairness and without detriment to the interests of either of them and that, if in the course of the performance of this Agreement, unfairness or detriment to either Party does or may result then the other shall use its reasonable endeavours to agree upon such action as may be necessary to remove the cause or causes of such unfairness or detriment.

3. INDEPENDENT CONTRACTORS

3.1 Although the parties have agreed to co-operate with each other in the manner described in this Agreement, the parties shall at all times be independent parties which are in business on their own account and neither party is an agent or partner of the other (within the meaning of the Partnership Act 1890) or authorised to assume or create any obligations or liabilities, express or implied, on behalf of or in the name of the other party.

3.2 The parties shall at all times be and remain responsible for the actions and omissions of persons employed or engaged by them in the performance of this Agreement.

3.3 The employees of each party shall at all times be and remain employees of that party and under the exclusive direction and control of that party.

4. CONTRACT MANAGEMENT

4.1 Both parties shall appoint an Authorised Representative who will be responsible for managing the relationship embodied by this Agreement.

4.2 The parties shall ensure that the Authorised Representatives meet with or speak to each other not less frequently than as described in Schedule 2 to this Agreement during the term of this Agreement to:

4.2.1 discuss the performance of Agreement.

4.2.2 deal with any requests for information;

4.2.3 deal with requests for any change in relation to this Agreement;

4.2.4 exchange information of mutual interest and identify any other projects of potential benefit arising out of or in connection with the Data Activities.

4.3 Neither party shall change its Authorised Representative without having first notified the other party in writing of the proposed change.

5. SPECIFIC OBLIGATIONS OF DATA PROCESSORS

5.1 Each Data Processor agrees that:

5.1.1 using all reasonable care and skill; and

5.1.2 adhering to the Standards and Policies; and

5.1.3 applying the Process

it shall use reasonable endeavours to perform the Data Activities.

5.2 Each Data Controller hereby confirms that their respective Data Processor is authorised to perform the obligations of the Data Controllers under this Agreement and that either Data Controller may liaise directly with either Data Processor in relation to any matter arising out of or in connection with the performance of the Data Activities hereunder.

5.3 Each Data Processor shall ensure that the minimum practicable number of Data Processor staff shall have access to the Identifiers and shall procure that they comply with the terms of this Agreement in respect of the same.

5.4 Each Data Processor agrees that they shall comply with the Process in respect of disclosure to third parties of the Output;

5.5 Where any timetable for the performance of the Data Activities is specified within this Agreement and its schedules, each Data Processor will endeavour to comply with the same but does not make any warranty or representation that it will be able to do so.

5.6 Each Data Processor agrees that they shall process the Identifiers and any other data provided to it by the Data Controllers only in accordance with this Agreement. Each Data Processor shall comply with any retention and/or destruction obligations with regard to the Identifiers and any other data provided by the Data Controllers, as described in the Process.

5.7 Each Data Processor agrees that in processing Personal Data, they shall comply with the provisions of the DPA in its capacity as Data Processor and act in accordance with the instructions of the Data Controller and after having implemented appropriate technical and organisational measures to protect the same against unauthorised or unlawful processing and accidental loss, destruction, damage, alteration or disclosure.

5.8 Each Data Processor agrees to comply with the supplemental terms and conditions (if any) set out in Schedule 1 to this Agreement.

6. SPECIFIC OBLIGATIONS OF THE DATA CONTROLLERS

6.1 Each Data Controller agrees that it will:

6.1.1 be responsible for ensuring all permission, licences and approvals it deems necessary are obtained for it to lawfully and ethically disclose the identifiers and any other data to its respective Data Processor and for that Data Processor to perform the Data Activities in the manner envisaged by this Agreement (e.g. Section 251 of the National Health Service Act 2006);

6.1.2 perform those tasks and activities for which it is responsible as described in and in accordance with the Process and, to the extent applicable to such tasks and activities, the Standards and Policies;

6.1.3 permit its respective Data Processor to identify itself and the Data Controller as participants in the project envisaged by the Purpose;

6.1.4 provide all necessary information, co-operation and assistance to its respective Data Processor and to any third party regulator or monitoring organisation arising out of or in connection with the performance of the Data Activities;

6.1.5 not unreasonably withhold or delay any approval or consents required of it in order for its respective Data Processor to perform the Data Activities.

6.2 Both Parties agree that they will comply with the provisions of the DPA in their respective capacities as Data Controllers.

6.3 Both Parties as Data Controllers agree to comply with the supplemental terms and conditions (if any) set out in Schedule 1 to this Agreement.

7. CHANGE CONTROL PROCEDURE

7.1 If either Party wishes to change this Agreement (including the terms of reference for the Project) for any reason, it may request such a change by notice in writing to the other. Requests should be addressed to the Authorised Representative.

7.2 If the Parties agree to implement a change, the details and impact of that change (including agreement as to scope of marketing arrangements, revised timetables of work and any costs) shall be recorded in writing and signed by both Parties. Neither party shall be under any obligation to effect any change until such time as written agreement has been reached in accordance with this Clause 7.

8. INTELLECTUAL PROPERTY RIGHTS OWNERSHIP AND GRANT OF LICENCE

8.1 DfE shall be and remain the owner (or licensee) of the intellectual property rights in all data, materials, information and documentation provided by DfE to SLaM regardless of whether the same were created or developed before or after entering into this Agreement. However, DfE hereby grants to SLaM, for the Term, a free of charge, non-exclusive licence to use the same solely for the purposes of enabling CDLS to perform the Data Activities in accordance with the Process.

8.2 SLaM shall be and remain the owner (or licensee) of the intellectual property rights in all data, materials, information and documentation provided by SLaM to DfE regardless of whether the same were created or developed before or after entering into this Agreement. However, SLaM hereby grants to DfE, for the Term, a free of charge, non-exclusive licence to use the same solely for the purposes of enabling DfE to perform the Data Activities in accordance with the Process.

9. DISPUTE RESOLUTION

9.1 The Parties shall attempt to resolve any dispute arising out of or relating to this Agreement through negotiations between the Authorised Representatives.

9.2 If the Authorised Representatives are unable, within fourteen (14) days of a dispute having been referred to them, to resolve such dispute, the matter shall be referred to appropriate senior management representatives who shall have authority to settle the same.

9.3 If the senior management representatives (described in clause 11.2) are unable, within fourteen (14) days of a dispute having been referred to them, to resolve such dispute, the Parties will attempt in good faith to resolve the dispute through an Alternative Dispute Resolution (ADR) procedure as recommended to the Parties by the Centre for Dispute Resolution.

9.4 If the matter has not been resolved by an ADR procedure within thirty (30) days of the initiation of that procedure, or if either party will not participate in an ADR procedure, the dispute shall be decided by the English courts.

10. TERMINATION

10.1 This Agreement shall continue for the Term unless terminated earlier in accordance with Clauses 10.2 or 10.3 below.

10.2 Either Party shall be entitled to terminate this Agreement for convenience by giving to the other not less than one (1) months' notice in writing.

10.3 Either Party shall be entitled to terminate this Agreement with immediate effect by notice in writing:

10.3.1 if the other has committed a material breach of its obligations under this Agreement and has failed to remedy that material breach within seven (7) days of having been requested to do so in writing; or

10.3.2 upon the other Party passing a resolution for winding-up (save for the purposes of amalgamation or reconstruction where the amalgamated or reconstructed company agrees to adhere to this Agreement) or suffering a winding-up order being made against it or going into administration; or

10.3.3 if a receiver or administrative receiver is appointed or an encumbrancer takes possession of the undertaking or assets (or any part thereof) of the other Party; or

10.3.4 if the other Party is unable to pay its debts (within the meaning of Section 123 of the Insolvency Act 1986 or any statutory re-enactment or modification thereof) or ceases to or threatens to cease to carry on its business or enters into a composition with its creditors.

10.4 Either Party shall be entitled to terminate this Agreement in the event of any breach by the other of Clause 15.2 or by anyone employed by it or acting on its behalf (whether with or without the knowledge of the other) or the commission of any offence under the Bribery Act 2010 by the other or by anyone employed by it or acting on its behalf.

10.5 Termination of this Agreement shall be without prejudice to any rights or obligations of either Party, which arose on or before its termination or which are expressed to arise upon or continue after termination.

11. LIMITATION OF LIABILITY

11.1 Neither Party excludes or limits its liability to the other Party in respect of fraudulent misrepresentation, breaches of confidentiality or for death or personal injury caused by its negligence.

11.2 The liability of either Party in respect of loss or damage to tangible, physical, property of the other Party caused by its negligence shall be limited to £1,000,000 per event or series of connected events.

11.3 The Parties hereby expressly agree that the other Party in complying with Schedule 1 is not responsible to and shall not bear any liability to any Data User in respect of the Data Activities and/or any Output nor shall Parties be responsible for the actions or omissions of any Data User arising out of or in connection with the performance of the Data Activities and/or relating to the use of any Output. The Parties shall do all such things as may be necessary to give effect to this clause and to ensure that such Data Users understand that any and all use by them of any Output is made at their own risk.

11.4 Subject to Clauses 11.1, 11.2 and 11.3, each Party's total aggregate liability to the other arising out of or in connection with this Agreement shall be limited to £1,000.

12. CONFIDENTIALITY AND FOIA

12.1 Neither Party shall, otherwise for the purposes envisaged by this Agreement, disclose to any person (other than with the written authority of the other) any confidential information concerning the information technology, patients, service users, clients, business operations, accounts, finance or contractual arrangements or other products, dealings, transactions or affairs of the other (including the other Party's business strategies or development plans) which may come or have already come to that Party's knowledge in the course of dealing with the other party (whether in connection with this Agreement or otherwise).

12.2 Nothing contained in this Clause 12 shall prevent either Party from disclosing that information:-

12.2.1 to any of its employees whose work requires the disclosure of that information and who have prior to the disclosure of that information agreed in writing to keep such information confidential;

12.2.2 to any government department or other authority, court or arbitrator having statutory authority or jurisdiction to require the disclosure of that information.

12.3 The obligations of confidentiality in this Agreement shall not apply to information which:

12.3.1 is already in the public domain or subsequently comes into the public domain otherwise than by breach of this Agreement, or

12.3.2 can be proven to have been known to the receiving Party at the time of acquiring it from the other Party; or

12.3.3 was disclosed or used with the prior written approval of the Party from whom it was lawfully acquired; or

12.3.4 was received from a third party which did not itself obtain it in confidence directly or indirectly from the supplying party; or

12.3.5 can be shown to have been independently developed by the receiving party; or

12.3.6 is trivial or obvious.

12.4 Both Parties acknowledge that the other is or may be subject to FOIA and the respective Codes of Practice on the Discharge of Public Authorities' Functions and on the Management of Records (which are issued under section 45 and 46 of the FOIA respectively) and the Environmental Information Regulations 2004 as may be amended, updated or replaced from time to time. Both parties will act in accordance with the FOIA, these Codes of Practice and these Regulations (and any other applicable codes of practice or guidance notified to the Contractor from time to time) to the extent that they apply to this Agreement.

12.5 Both Parties agree that:

12.5.1 the provisions of Clause 12.1 are subject to the respective obligations and commitments of the parties under FOIA and both the respective Codes of Practice on the Discharge of Public Authorities' Functions and on the Management of Records (which are issued under section 45 and 46 of the FOIA respectively) and the Environmental Information Regulations 2004;

12.5.2 the decision on whether any exemption applies to a request for disclosure of recorded information is a decision solely for the applicable Party;

12.5.3 where a party is managing a FOIA request, the other shall co-operate with it and shall respond within five (5) working days of any request by it for assistance in determining how to respond to a request for disclosure.

12.5.4 Both Parties shall and shall procure that its sub-contractors shall:

(a) transfer any request under FOIA for information, as defined under section 8 of the FOIA, to the other as soon as practicable after receipt and in any event within five (5) working days of receiving a request for information;

(b) provide the other with a copy of all information in its possession or power in the form that the other requires within five (5) working days of the other requesting that information and provide all necessary assistance as reasonably requested by the other to enable the other to respond to a request for information within the time for compliance set out in section 10 of FOIA.

13. NON-SOLICITATION OF PERSONNEL

13.1 Neither Party shall, without the prior written consent of the other Party, actively initiate recruitment of any of the employees of the other who are engaged in the performance of the Data Service, during the Term or for a period of one year after termination of this Agreement.

13.2 In recognition of the value of the personnel and the inconvenience which would be caused as a result of a breach of Clause 13.1, each Party agrees that, if it does breach Clause 13.1, it shall pay to the other Party:

13.2.1 in respect of an employee of the other Party – an amount which is equivalent to that employee's gross salary over the six (6) months immediately preceding the date of termination of his/her employment with the other party;

13.2.2 in respect of any other person – an amount which is equivalent to the gross revenue generated (from all sources) by that person during the six (6) months immediately preceding the date of termination of his/her engagement with the other party.

13.3 The Parties hereby expressly agree that the sums referred to in Clause 13.2 represent a genuine pre-estimate of the loss likely to be suffered in the circumstances described at Clause 13.1.

14. NOTICES

14.1 Any notice given under this Agreement must be given in writing and sent or delivered by hand, post, or email to the other party at the address stated in Schedule 2 to this Agreement (or any other address notified for this purpose by that Party) provided that any:

14.1.1 notice delivered by hand shall be deemed to have been given when deposited at the appropriate address;

14.1.2 notice sent by post shall be deemed to have been given forty eight (48) hours after a first class registered letter is posted to the appropriate address; and

14.1.3 notice sent by email shall carry a marker which identifies when the email has been received and shall be deemed to have been given when electronic confirmation of receipt is indicated.

15. GENERAL

15.1 Both Parties shall in all matters arising out of or in connection with the performance of the Agreement comply with the law and with all orders, regulations and bye-laws made with statutory authority by Government Departments or by local or other authorities that are applicable to the Agreement, from time to time.

15.2 Both Parties shall be obliged to immediately notify the other in writing, if it becomes aware of any legislation, rules or guidance which might impact upon the lawful performance of this Agreement.

15.3 Neither Party shall:

15.3.1 offer or give or agree to give any servant of the other any gift or consideration of any kind as an inducement or reward for (a) doing or forbearing to do or (b) having done or forborne to do any act in relation to the obtaining or performance of this Agreement or (ci) showing or forbearing to show favour or disfavour to any person in relation to this Agreement; nor

15.3.2 enter into this Agreement if any commission has been paid or agreed to be paid (a) by it or (b) (on its behalf or to its knowledge) to any person employed by or in the service of the other party, unless before the Agreement has been entered into, particulars of the same have been disclosed in writing to the Authorised Representative of the other party.

15.4 The Parties hereby expressly agree that any person who is not a party to this Agreement shall have no right to enforce any term of this Agreement against either of the Parties pursuant to the Contracts (Rights of Third Parties) Act 1999.

15.5 No failure, delay or indulgence on the part of either party in exercising any power or right under this Agreement shall operate as a waiver of such power or right.

15.6 If any provision of this Agreement shall be held by a court of competent jurisdiction to be invalid or voidable such provision shall be struck out and the remainder shall stand in full force or effect.

15.7 Neither Party may assign or novate this Agreement or any of its rights and obligations thereunder without the prior written consent of the other.

15.8 Neither Party shall be liable for delay or failure to perform any of its obligations under the Agreement insofar as the performance of such obligation is prevented by any event beyond its reasonable control (*an event of force majeure*).

15.9 Rights and obligations of the parties which have accrued or which shall accrue shall survive termination of this Agreement insofar as survival may be construed from the relevant clauses in the context of such termination. The obligations of confidentiality under this Agreement shall survive in perpetuity.

15.10 The Parties hereby expressly agree that money damages may not be a sufficient remedy for any breach of this Agreement and that the party not in breach shall be entitled to equitable relief, including injunction and specific performance. Such remedies shall be in addition to all other remedies available at law and in equity.

16. ENTIRE AGREEMENT AND LAW

16.1 This Agreement constitutes the entire agreement between the parties with respect to the subject matter contained therein. All prior agreements, representations,

statements, negotiations, understandings and undertakings either written or oral are, unless made fraudulently, superseded hereby and the parties hereby acknowledge that they have not placed any reliance on any representation made but not embodied in those documents.

16.2 No change to this Agreement or any waiver of any of the terms hereof shall be valid unless made in writing and signed by the duly authorised representatives of both parties.

16.3 This Agreement shall be subject to English law and the Parties agree to submit to the exclusive jurisdiction of the English courts.

17. TERM OF AGREEMENT

17.1 Subject to clause 17.2, this Agreement shall expire on the date specified in Schedule 1, unless terminated earlier in accordance with clause 10.

17.2 The Parties may by mutual agreement in writing agree to extend the term of this Agreement beyond the date set out in Schedule 1.

SCHEDULE 1
PROJECT DESCRIPTION

South London & Maudsley (SLAM) NHS Foundation Trust proposed data linkage with National Pupil Database (NPD)

The South London & Maudsley (SLAM) NHS Trust collects and manages data on their patients. Data include personal, demographic, diagnostic and treatment information. The data are managed by a Confidential Data Linkage Service (CDLS) and stored on a Case Record Interactive System (CRIS).

The Trust collect data on young people aged 5-17 who access Child and Adolescent Mental Health Services (CAMHs). The Trust is keen to match these data for a cohort of young people (approximately 35,000 records) with DfE National Pupil Data (from the NPD). This will allow the Trust to investigate: the educational histories of young people accessing their services; the effectiveness of interventions to support later educational outcomes. The findings will be useful to promote school attendance and attainment for young people with mental health problems.

The proposed process is as follows:

1. SLAM researchers identify a cohort of young patients. This is drawn from an extract of the main database, holding anonymised data with a local identifier. This is sent to CDLS staff.
2. CDLS staff use the local identifiers to access personal level data. A dataset of patients will be drawn, including: name, date of birth, gender, postcode and local identifier. These data will be sent to DfE securely.
3. DfE will match these data to the NPD for agreed variables/years. In addition, a comparison sample of young people who are not SLAM patients but live in the same area will be identified and similar data drawn.
4. DfE will return data to CDLS, without personal data but including the local identifier and denoting patient/non patient. DfE will destroy CDLS data.
5. CDLS will store data securely, match data to assessment and treatment information, and draw anonymised samples in response to requests from SLAM researchers.
6. DfE will have opportunity to comment on: proposals for data for analysis; reports of findings.

At all times SLAM researchers analyse anonymised data sets. Patients are not explicitly asked by medical staff for consent to link their data to other databases. However, the system has received full ethical approval, formal Executive approval and SLAM Trust Caldicott approval. The latter provides clear guidance on the protection and use of patient information, while noting the need to comply with the data protection act.

CDLS has successfully completed several large scale secure data linkages with primary and secondary health care datasets. CDLS gained National Information Governance Board approval for these data linkages; the approval is an S251 approval and provides consent to the linkage on behalf of patients. CDLS actively communicate their data activity to patients; it is not considered reasonable/feasible/practical to seek consent directly from patients.

SCHEDULE 1
PROJECT DESCRIPTION

The rationale for this project has been approved by the Health Research Authority (HRA) Confidentiality Advisory Group (CAG) through s.251 approval. As such, the written application approved through s.251 constitutes Schedule 1 (or 'the Process' etc. as referenced in the body of the doc) to this agreement.

Appendix 1.1.1-6	Letter acknowledging s.251 SLaM application final approval by the HRA CAG.
Appendix 1.2.1-6	Letter acknowledging s.251 SLaM application provisional approval by the HRA CAG.
Appendix 1.3.1-10	SLaM Response to HRA provisional approval.
Appendix 1.4.1-30	SLaM application to HRA CAG for s.251 approval. Part B: Section 8 of this document (Appendix 1.4.24) provides the data flow of this project. For clarity the role and responsibilities of Data Processor we be assumed by; • DfE from the point of receiving the BRC Identifiers table from SLaM in Step 3 until the CDLS confirms receipt of the BRC cases and control NPD table in Step 5. • SLaM from the point of CDLS confirming receipt of the BRC cases and controls NPD table from DfE in Step 5 until completion of Step 8.
Appendix 2.1-3	Schedule to the Agreement for the supply of NPD Data This defines the content and structure of NPD data that will be transferred between Parties.

SCHEDULE 2
ADMINISTRATION

1. **AUTHORISED REPRESENTATIVES**

   1.1 For SLaM – Mr Stephen Docherty

   1.2 For DfE – Mr Richard Lumley

2. **ADDRESS FOR SERVICE OF NOTICES**

   2.1 For SLaM – CDLS SL&M Biomedical Research Centre Nucleus, Maudsley Site, Ground Floor, Mapother House, De Crespigny Park, Denmark Hill, London, SE5 8AF

   2.2 For DfE –National Pupil Database & Transparency Unit, Education Data Division, Department for Education, Mowden Hall, Staindrop rd, Darlington, DL3 9BG

3. **FREQUENCY OF CONTRACT MANAGEMENT MEETINGS** (see Clause 4.2):

   3.1 Contract Management Meetings not to be scheduled in advance. They will be arranged as and when needed by mutual agreement between the authorised representatives and CDLS. In absence of any arranged meeting following the linkage of CRIS and NPD, a summary of linkage project activity will be provided by CDLS to Project Management team and key stakeholders every 3 months.

SCHEDULE 3
PART A
CDLS STANDARDS

1. Introduction

South London and Maudsley Foundation Trust (SLaM) shall use all reasonable endeavours to comply with these Standards which serve to demonstrate its understanding of the Information Governance (IG) risks and the security and confidentiality needs of processing data within the Clinical Data Linkage Service (CDLS).

2. Service Details

2.1 The name of the service will be the SLaM IG Clinical Data Linkage Service (CDLS).

2.2 The service's responsible owner will be Stephen Docherty, Chief Information Officer, SLaM.

2.3 The service's Caldicott Guardian will be Dr Dele Olajide, SLaM Caldicott Guardian.

2.4 The Data Controller/s will be the Party/ies providing databases to CDLS for the purpose of either (i) enabling such data to be linked with at least one other database and holding and/or hosting such data, or (ii) for the purpose of holding and/or hosting data which has already been so linked.

(NB. The process of linking the datasets may be performed by CDLS or externally to CDLS prior to transfer to CDLS, as indicated in Schedule 1)

3. CDLS Security

3.1 Service Security: The SLaM ICT estate and ICT network service are governed by the SLaM ICT Security Policy.

3.2 Security Manager: The service's responsible security manager will be SLaM's Deputy Director of ICT (Ricky Mackennon), who will be responsible for:

3.2.1 Defining local access control policies on specific local assets,

3.2.2 Implementing and enforcing local access control policies,

3.2.3 Leading the SLaM ICT Support Team

3.2.4 Working with local developers on deployment and management of security solutions providing access to assets,

3.2.5 Security sign-off/accreditation of systems giving access to assets,

3.2.6 Continuous monitoring and evaluation of CDLS assets and their interfaces,

3.2.7 Monitoring and reporting actual or potential security breaches to the Director of ICT and the SLaM Caldicott Guardian,

3.2.8 Audit of detailed procedures to ensure that service security is maintained,

3.2.9 Ensuring that the service operates in accordance with the SLaM ICT Security Policy at all times

3.2.10 Ensuring that staff are aware and receive training in compliance through the SLaM induction and service training programme(s) in accordance with the ICT Training Schedule

3.3 Security countermeasures:

3.3.1 Physical security measures:

(a) All service equipment will be located within SLaM ICT premises in secure areas with restricted access. Central systems areas such as a telecommunications hub or a server farm will be a high security area with an incorporated entry restriction and detection system. All data resources including servers which store confidential data will be located in a secure area with restricted access.

(b) No portable media that does not meet the NHS Connecting for Health encryption standard will be utilised.

(c) All terminals and devices used for the data linkage service will have password boot and encryption implemented to the NHS Connecting for Health IG standard.

(d) Equipment and data is not permitted to be taken off site without formal approval.

(e) Ports to be restricted to those required e.g. all ports (FTP, Telnet) not required to utilise or monitor the application/server to be disabled.

(f) Any device used for the data linkage service will be supported with uninterruptible backup power units. The backup units will complement any emergency generator power system to minimise loss of data and non-availability of systems. Uninterrupted Power Supply (UPS) includes the controlled shutdown of servers and services.

3.3.2 Access Control and Privilege Management:

(a) Access to service facilities will be limited to specific staff, whose job function requires access.

(b) All access to personal identifiable information is enabled via individual user accounts, which are subject to SLaM ICT security protocols and audit procedures.

(c) All staff with access to personal identifiable information are made aware of their responsibilities in the SLaM Confidentiality Policy and the ICT Security Policy.

(d) All data processed by CDLS will be stored on SLaM Servers with security permissions set for users with a legitimate right of access.

3.3.3 Network Security Measures:

(a) All information transfers will use NHSmail or the SLaM Secure Electronic File Transfer tool, which provide the level of encryption required by the NHS Connecting for Health or the equivalent secure data transfer portal via the Department for Education Key to Success website.

(b) All networked applications will utilise a security model with access controlled user accounts utilising user identity authentication and the SLaM active directory linkage via the use of domain password authentication (single sign on). Regular password changes will be mandatory every 30 days.

(c) The network will be enabled to only have required services and or protocol open, including the use of Access control lists.

(d) CDLS will adhere to the NHS data security standards set out in the Code of Connection.

3.3.4 Other authentication, certification, security testing, and audit:

(a) All service related devices will be installed in accordance with the manufacturers' and SLaM ICT department recommendations where applicable. Security and environmental controls will be implemented to SLaM standards when deemed applicable by the ICT Department.

(b) If there is a need for any disk / media utilised for the data linkage service to be removed from SLaM premises by a third party, the third party will be required to sign a declaration of confidentiality and information security with SLaM.

(c) SLaM has been undertaking a programme of activities to align its security management practices with those set out in ISO27001 Information Security Standard.

(d) SLaM undertakes the annual DoH IG Toolkit assessment to demonstrate compliance with key information governance, security, data protection and confidentiality, secondary uses, clinical and corporate records standards.

(e) All files received and generated during the course of the function of the data linkage service will be checked for malicious code before use or transmission.

(f) All devices utilised for CDLS will be kept under regular review for the purposes of ensuring that, where necessary, the latest SLaM supported software and/or patch is installed. All patch and updated version installation will be subject to the SLaM ICT Change Control procedure.

(g) Server build configuration documentation will be reviewed periodically or as and when a change is required.

(h) Access to personal identifiable information is monitored using internal system controls as well as an annual programme of assurance audits.

4. Service Management:

4.1 All SLaM systems are implemented and maintained in-house by the SLaM ICT Department as described in the documentation listed in section 9.2 of this document.

4.2 On occasion, CLDS may transfer data to, and receive data from, an external trusted third party e.g. The Health and Social Care Information Centre (HSCIC) in specific accordance with the data linkage process described in Schedule 1.

5. Service Design

5.1 The infrastructure at SLaM is provisioned as a vendor certifiable class leading virtual environment with Cisco UCS deployed as the blade infrastructure. CDLS uses desktop PCs owned and operated by SLaM. SLaM PCs are Dell Optiplex series (currently Optiplex 780).

Access to all CDLS PCs will be password protected using authenticated SLaM network user name and password.

All reasonable security steps have been taken to protect CDLS data from being accessed other than from inside the SLaM firewall. All CDLS PCs are on the SLaM Local Area Network (LAN) which is secure firewall protected. No standalone devices are used. No data is permitted to be stored on the local drives and security measures designed to protect data from being saved outside the firewall are in place.

5.2 The SLaM network conforms to the Information Governance Assurance Statement for Organisations that use the NHS National Network (N3) and other national applications like NHS Mail and Choose and Book. This process includes terms and conditions for use of NHS CFH systems and services including the N3, in order to preserve the integrity of those systems and services against malicious code, spam, hacking etc. Compliance with the requirements in the IG Toolkit allows the Trust to accept the IG Assurance Statement. This is re-confirmed each year by the submission of the annual IG Toolkit Assessment.

5.3 All PCs are networked to ensure they receive the necessary system updates, including malicious code updates that are delivered via the network, maintaining the security of their operating systems. Trend® Anti-Virus is used throughout SLaM and is centrally managed for the update and distribution of virus update files. Cisco ASA firewalls are used and arranged in clusters for failover. These new firewalls help to protect SLaM from internet borne threats and also from unknown traffic on the National N3 network. A centralised firewall reporting solution will be linked to the SLaM Enterprise Monitoring solution to bring together all of the systems information into a notional dashboard.

6. Operational Processes

6.1 Patient Identifiable Information (PII): Patient identifiable data and other sensitive data will be collected, used, processed and disposed of in accordance with the following SLaM policies (also *see section* 9.2.1-9.2.6):

6.2 Data Storage:

6.2.1 Format: The service will fully use on-line electronic means and no paperwork will be generated other than print outs of reports and activities.

6.2.2 Location: All data will be stored electronically on the SLaM network in accordance with the policies listed above

6.2.3 Anonymisation: CDLS will be capable of anonymisation, pseudonymisation and masking of PII/sensitive data. Anonymisation/masking of data within the secure CDLS, and 'output' will be anonymised according to specific requirements detailed in Schedule 1.

6.2.4 Encryption Standards: SLaM uses a variety of tools and solutions to manage the encryption of data. Data interchange is encrypted/decrypted to AS-256 password based encryption using a selection of tools. 7zip is a tool commonly used for this purpose. Additionally, SLaM is able to encrypt to FIPS 140-2 as required. This is validated for the Libgcrypt module that can be utilised either with GNUPG or certain enterprise Linux installation.

6.3 Data Processing:

6.3.1 CDLS desktop PCs will be used to access and process extracted data subsets. A CDLS approved user password enables access to these computers.

6.3.2 Data accessed and processed is stored only on the encrypted CDLS shared drive. No data is ever cached.

6.3.3 The SLaM Virtual Private Network (VPN) terminal server provides remote access to the data. Authentication and encryption is used to logon to the VPN service with limited session times. Users are advised that physical security precautions must be taken when working remotely, including avoiding use in public spaces, locking the screen when leaving computers/mobile devices unattended, being vigilant and aware of environmental threats to information security.

6.3.4 Transmission: All data transfers will be utilised by online means that comply with the NHS Connecting for Health encryption standard, including NHSmail

encrypted portable media. Postal services will not be utilised unless authorised by the Data Controller.

6.3.5 Use of unencrypted USB memory sticks and portable media such as CDs and DVDs has been disabled for the storage and transfer of person identifiable information. A SLaM issued standard memory stick is required to be used as these are password protected and encrypted.

6.4 Service's Authorised User: Only CDLS staff will have access to the databases. Data Users will be granted access to these databases in accordance with Schedule 1.

6.5 Data Disposal: When the data used has completed its purpose, is no longer required and has been retained in accordance with Schedule 1, it will be disposed of in accordance with the SLaM Confidential Waste Procedure and the ICT Security Policy. Data from servers will be wiped clean using the Darik's Boot and Nuke Program (DBAN) which is designed to protect against deleted data being retrieved.

7. Service Audit

7.1 Compliance with the SLaM ICT Security Policy and other IG policies is reviewed as part of the SLaM IG Assurance Programme. The focus of the annual assurance programme is varied in order to improve compliance and provide relevant assurance against the standards that make up the DoH IG Toolkit for Mental Health trusts.

7.2 In addition to audits targeted at key areas and functions of the SLaM ICT Department, the systems are subject to an on-going audit to identify potential inappropriate access, misuse of information and security risks. Regular reports are presented to the SLaM ICT Security and Caldicott Committees to ensure progress on the recommended actions are effectively monitored, gaps and weaknesses are identified and remedial actions are taken.

7.3 All information and systems related risks are assessed monthly and any changes to the impact and the likelihood of risks and progress of action plans outlined to mitigate the risks are reported to the Chief Executive of SLaM as part of the monthly performance review process. The ICT Risk Register is part of the SLaM Corporate Risk Register.

8. Service Protection

8.1 System Recovery Mechanisms

8.1.1 The system will backed up daily on a designated server and will be stored at two data warehouses at different locations, which are both directly managed and maintained by SLaM. Access to either data warehouse is restricted to authorised staff.

8.1.2 The purpose of such back-ups is to facilitate restoration of the database system and associated servers in the event of an electronic system failure and for major incidents, e.g. fires, the data and systems may then be re-established using data kept offsite.

8.1.3 SLaM servers and systems each have disaster recovery plans that are reviewed annually.

8.2 Business Continuity plan

8.2.1 In the event of serious disruption or total system failure, business continuity shall be provided according to the ICT Disaster Recovery and Business Continuity Plan.

8.3 Security or Confidentiality breach response procedure

8.3.1 Any incident that involves the loss of personal information, loss of ICT equipment, intentional or unintentional disclosure of personal identifiable information outside the legal framework of the Data Protection Act, the Caldicott Guidelines and SLaM IG policies is reported using the online

Datixweb incident reporting tool. All such incidents are automatically reported to the SLaM Caldicott Guardian, the SIRO and the Head of Information Governance.

8.3.2 Once notification of serious untoward incidents is received, the Head of IG will request a fact finder report from the relevant service manager and escalate the incident to the relevant Service Director and the relevant Executive. Once the fact finder report is finalised, the incident will be reported externally to Monitor, and the Information Commissioner's Office.

8.3.3 Information incidents are reviewed by the Head of IG every month and are regularly reported to the Caldicott and/or the ICT Security Committees using the classification endorsed by the Department of Health.

8.3.4 Any concerns of inappropriate access to the data processed by CDLS will be treated as a serious incident.

8.3.5 SLaM follows the Department of Health Checklist for Reporting, Managing and Investigating IG Serious Untoward Incidents.

8.3.6 For further information, staff can refer to the SLaM Information Risk, Incident and Forensic Readiness Policy.

9. CDLS Standards Document Ownership

9.1 This Standards document will be the responsibility of the SLaM Head of IG and will be reviewed annually.

9.2 Links from the SLaM IG intranet site for several documents mentioned in this document are provided below:

9.2.1 The SLaM ICT Security Policy

(file://slam/resources/ict/policies_and_procedures/ICT%20Security%20Policy%20v5.1%20August%202013.pdf)

9.2.2 The SLaM Confidentiality Policy

(http://sites.intranet.slam.nhs.uk/Policies/Trust%20Corporate%20policies/Confidentiality%20Policy%20v6%20August%202013.pdf)

9.2.3 The KHP Information Sharing Policy

(http://sites.intranet.slam.nhs.uk/Policies/Trust%20Corporate%20policies/KHP%20Information%20sharing%20policy%20v2%20June%202013.pdf)

9.2.4 The SLaM Information Governance Policy

(file://slam/resources/ICT/policies_and_procedures/Information%20Governance%20Policy%20v5%201%20June%202013.pdf)

9.2.5 The SLaM Information Risk, Incident and Forensic Readiness Policy

(http://sites.intranet.slam.nhs.uk/Policies/Trust%20Corporate%20policies/Information%20Risk,%20Incident%20and%20Forensic%20Readiness%20Policy,%20v1%20-%20April%202011.pdf)

9.2.6 The SLaM ICT Disaster Recovery and Business Continuity Plan (http://sites.intranet.slam.nhs.uk/ICT/itservices/IT_Internal/Business%20Continuity%20Planning/Slam%20Disaster%20Recovery%20Plan%20V0.9.pdf)

9.3 Other associated documents include the full range of IG policies outlined in the SLaM IG Management Framework are shown in figure 1.

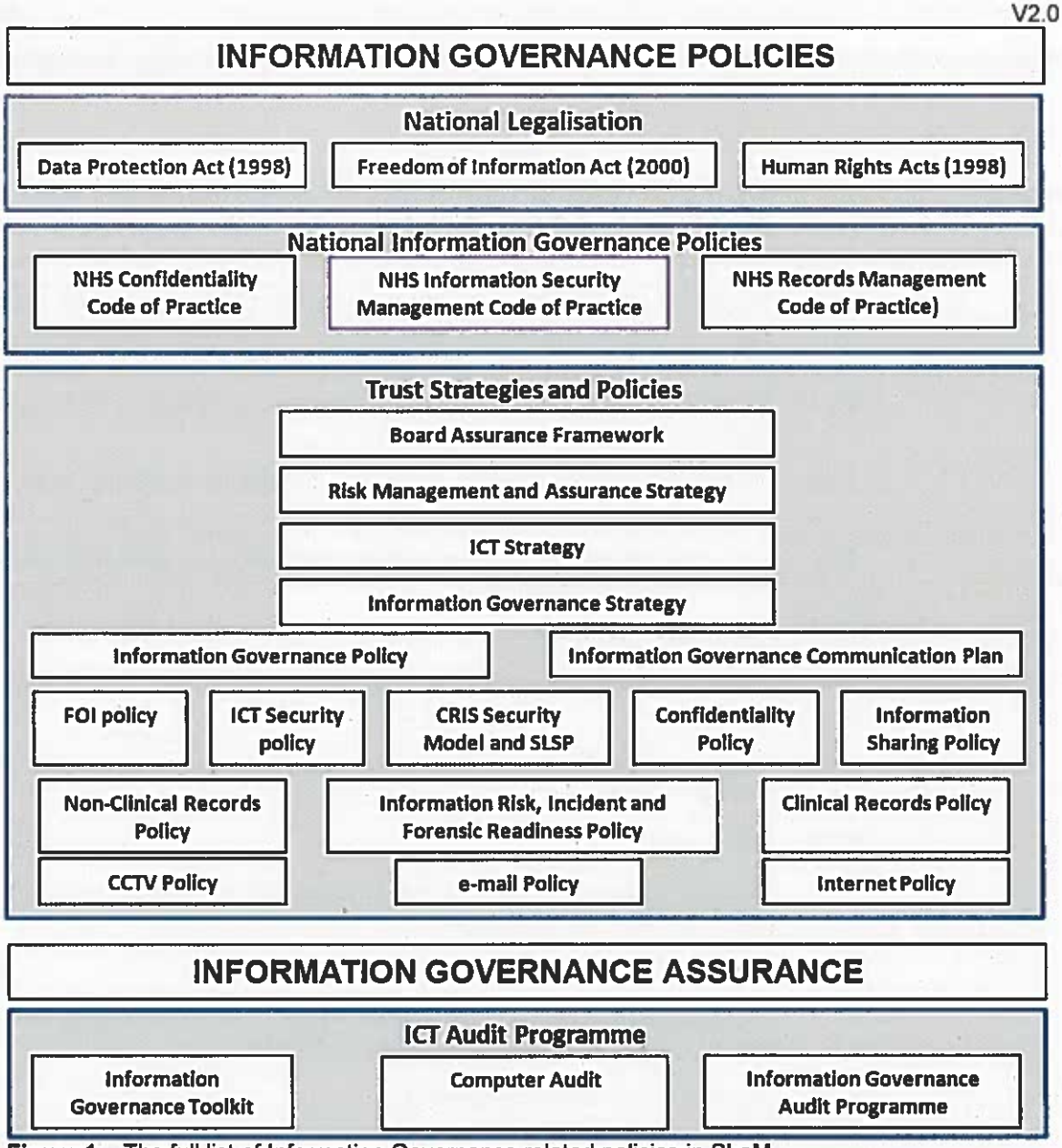


Figure 1 – The full list of Information Governance related policies in SLaM

10. Data Protection Registration

10.1 SLaM Data Protection Registration (Ref No: Z6032780) covers the purposes of analysis for the classes of data requested. The details of the registration is available at www.ico.gov.uk

SCHEDULE 3
PART B
GENERALLY APPLICABLE STANDARDS AND POLICIES

1. Personal Data and medical records (Defined terms shall have the meanings set out in the DPA):

1.1 Both parties shall comply with the Data Protection Act 1998 and any other applicable data protection legislation. In particular, if acting as Data Processor, each party agrees to comply with the obligations placed on the other (as Data Controller) by the seventh data protection principle ("the Seventh Principle") set out in the DPA, namely:

1.1.1 to maintain technical and organisational security measures sufficient to comply at least with the obligations imposed on the other by the Seventh Principle;

1.1.2 only to process Personal Data for and on behalf of the other, in accordance with the instructions of the other and for the purpose of performance of the Agreement and to ensure compliance with the DPA;

1.1.3 to allow the other to audit the party's compliance with the requirements of this paragraph 1.1 on reasonable notice and/or to provide the other with evidence of its compliance with the obligations set out in this paragraph 1.1

1.2 Each party agrees that it shall, in the event of a data security breach discovered by it, it shall:

1.2.1 within twenty four (24) hours of becoming aware of such breach, notify the other and immediately take all reasonable steps to mitigate and remedy such breach and prevent the re-occurrence of such breach, including taking such steps as may reasonably be required by the other;

1.2.2 provide the other with all such information as it reasonably requires in relation to the breach to enable it to assess the nature and extent of the breach and any action which the other may be required to take in response to such breach; and

1.2.3 except to the extent required by law and/or any regulatory body, not disclose the fact of such breach to any third party, without the prior written consent of the other.

1.3 Neither party shall store or process Personal Data for the purposes of this Agreement at sites outside the United Kingdom.

1.4 Both parties shall take steps to ensure the reliability of any of its personnel who will have access to Personal Data.

1.5 Both parties shall and shall procure that all persons acting on its behalf who are processing Personal Data comply at all times with the data protection legislation and shall not act or omit to act in such a way as to cause the other to breach any of its applicable obligations under the data protection legislation.

1.6 Both parties agree to use all reasonable efforts to assist each other to comply with the DPA. For the avoidance of doubt, this includes each party providing the other with reasonable assistance in complying with subject access requests served on the other under Section 7 of the DPA and consulting with the other prior to the disclosure by that party of any Personal data in relation to such requests.

1.7 For the avoidance of doubt, failure to comply with this paragraph (1.) may be treated by the other as constituting a material breach of contract entitling the other party to terminate this Agreement.

2. Data Encryption Standards and Policies

2.1 Both parties shall ensure that they transfer to each other only such data as is essential (having regard to the Purpose) to enable the provision of the Data Service.

2.2 Both parties shall ensure, in respect of all data which is transferred electronically (including but not limited to data transferred over wired or wireless networks, held on laptops, CDs, memory sticks and/or other moveable devices and/or media), that they shall apply the data encryption Standards and Policies described in section 9.2 of Schedule 3a to this Agreement:

3. Compliance with Section 251 of the National Health Service Act 2006 (s.251) conditions:

3.1 Both parties shall at all times comply with all agreements, regulations and directions made pursuant to s.251 (as amended).

4. SLaM ICT security policies and arrangements

4.1 SLaM shall ensure that, to the extent applicable to the Data Activities, CDLS complies with all SLaM ICT security policies and procedures, as the same are amended or updated from time to time.

Signed for and on behalf of South London and Maudsley NHS Foundation Trust:

Signed..... .....

Name.....Stephen Docherty.....

Position.....Chief Information Officer.....

Date.....13th January 2015

Signed for and on behalf of Department for Education:

Signed.....Elizabeth Tadd.....

Name.....ELIZABETH TADD.....

Position.....HEAD OF BUSINESS IMPROVEMENT.....

Date.....01/04/15.....